2010

Development and Implementation of a Tissue Specific MicroRNA Prediction Tool for Identifying Targets of the Tumor Suppressor microRNA 17-3p

William Budd
Virginia Commonwealth University

William Thomas Budd, BS Bioinformatics

This is to certify that the thesis prepared by William T. Budd entitled *Development and Implementation of a Tissue Specific MicroRNA Prediction Tool for Identifying Targets of the Tumor Suppressor microRNA 17-3p* has been approved by his/her committee as satisfactory completion of the thesis requirement for the degree of Master of Science in Bioinformatics

_____
Zendra Zehner, PhD Department of Biochemistry and Molecular Biology

_____
Joy Ware, PhD Department of Pathology

_____
Lemont Kier, PhD Center for the Study of Bilogical Complexity

_____
Nihar Sheth, MS Center for the Study of Bilogical Complexity

_____
Greg Buck, PhD Director Center for the Study of Biological Complexity

_____
Thomas Huff, PhD Vice Provost of Life Science

_____
Douglas Boudinot, PhD Dean of Graduate Studies

May 12, 2010

Development and Implementation of a Tissue Specific MicroRNA Prediction Tool for Identifying Targets of the Tumor Suppressor microRNA 17-3p

A thesis submitted in partial fulfillment of the requirements for the degree of Masters of Science at Virginia Commonwealth University.

By

William T. Budd

BS Virginia Commonwealth University 2009

Director: Zendra Zehner, PhD

Department of Biochemistry and Molecular Biology

Virginia Commonwealth University

Richmond, Virginia

April 2010

Acknowledgements

The author wishes to acknowledge the extreme patience demonstrated by his wife and children during the pursuit of his graduate degree. Without their understanding and compassion, the completion of this body of work would not have been possible. This thesis is dedicated in part to my wonderful wife Lori. I would like to thank her for her help, support and continued encouragement. I would also like to dedicate this work to my grandmother, Iris T. Paris. Her guidance, compassion and support made me the man I am today.

Several members of my committee and lab offered expertise and advice that were invaluable to this work. I graciously thank them for their efforts and acknowledge their valuable insight. A special thanks goes out to Dr. Zendra Zehner for allowing me the opportunity to work with her on this very important project. Her leadership, support and guidance were very valuable to the completion of this project. Xueping Zhang, a post-doctoral fellow in the Zehner lab, offered advice and expertise on various experimental methods and protocols used in this thesis. Without her help and expertise, wet lab confirmation of our hypotheses would not have been possible. Nihar Sheth guided the computational design of the program. He provided a sounding board for ideas, helped to troubleshoot the computational programs and was invaluable in the design of the web interface.

TABLE OF CONTENTS

List of Tables

Table Page

List of Figures

List of Abbreviations

BPH- Benign Prostatic Hypertrophy

Contig- Contiguous Sequence

DNA- Deoxyribonucleic Acid

EMBL- European Molecular Biology Laboratory

EST- Expressed Sequence Tags

FDA- Food and Drug Administration

FTP- File Transfer Protocol

IGF1R- Insulin Growth Factor Receptor 1

Kcal/ mol- Kilocalories per Mole

MAPI- MicroRNA Annotation and Prediction Interface

miR – microRNA

miRNA- microRNA

miTG – microRNA Target Gene Score

MRE- microRNA Recognition Element

mRNA- Messenger RNA

NCBI- National Center for Biotechnology Information

NTC – Non-targeting Control

oncomiR- Oncogenic microRNA

PBS- Phosphate Buffered Saline

PERL- Pattern Expression and Reporting Language

PSA- Prostate Specific Antigen

Refseq- Reference Sequence

RNA- Ribonucleic Acid

RNA pol II- RNA polymerase II

RNA pol III- RNA polymerase III

RPM- Revolutions per Minute

RPMI 1640 – Roswell Park Memorial Institute Growth Medium

SDS- Sodium Dodecyl Sulfate

shRNA- Small Hairpin RNA

TBST- Tris Buffered Saline Tween 20

UTR- Untranslated Region

YES1- Yamaguchi Sarcoma Virus Oncogene Homolog

Abstract

DEVELOPMENT AND IMPLEMENTATION OF A TISSUE SPECIFIC MICRORNA PREDICTION TOOL FOR IDENTIFYING TARGETS OF THE TUMOR SUPPRESSOR MICRORNA-17-3P

A unique computational approach was undertaken to identify targets of miR-17-3p that impart an oncogenic potential to the cells of the prostate. Utilizing this approach, we identified insulin growth factor receptor 1 (*IGF1R*) as a potential target of miR-17-3p. *IGF1R* imparts an oncogenic approach to the cells by helping cells escape apoptosis, become hypertrophic and increase the production of extracellular proteases that allow cells to detach from neighbors.

The regulation of insulin growth factor receptor 1 by human microRNA-17-3p was evaluated using a western blot analysis of prostate cancer cell lines. Protein levels were compared in a cell line that expressed a non-targeting control RNA and a cell line that expressed microRNA-17-3p. The cell line that expressed the non-targeting control had significantly higher levels of IGF1R protein than the cell line expressing more of the active microRNA. Based on this experiment, it appears that microRNA-17-3p might regulate the insulin growth factor receptor 1.

Chapter 1:

Introduction

The Prostate Gland

The male prostate gland is a walnut shaped exocrine gland about four centimeters in diameter and has a mass of approximately 20 grams when fully mature. The prostate is located inferior to the urinary bladder, and anterior to the rectum. Passing through the prostate is the prostatic urethra. In boys, the prostate gland is very small and begins to hypertrophy as they approach adolescence and reaches its mature size shortly after puberty. Under normal circumstances, the prostate gland ceases to grow. In half of all men, when they reach an approximate age of fifty, the gland begins to hypertrophy again. This results in a condition known as benign prostatic hypertrophy (BPH). During the progression of BPH, the prostate gland begins to compress the urethra and causes great difficulty in urination. Many men over the age of fifty suffer from increased frequency of urination, hesitancy and urinary incontinence.

This prostate gland serves several important functions in the male genitourinary system. Secretions of the prostate account for approximately two thirds of the fluid content of semen. They include an alkalinic substance that counteracts the acidity of the vagina allowing sperm cells to survive the harsh environment that they encounter on their journey to the mature ovum. The prostate gland is composed in part of smooth muscle that contracts when stimulated increasing the velocity of ejaculate through the urethra. During ejaculation the prostate gland will contract and block the flow of urine into the urethra during ejaculation by closing off the portion of the urethra coming from the urinary bladder.

Incidence of Prostate Cancer

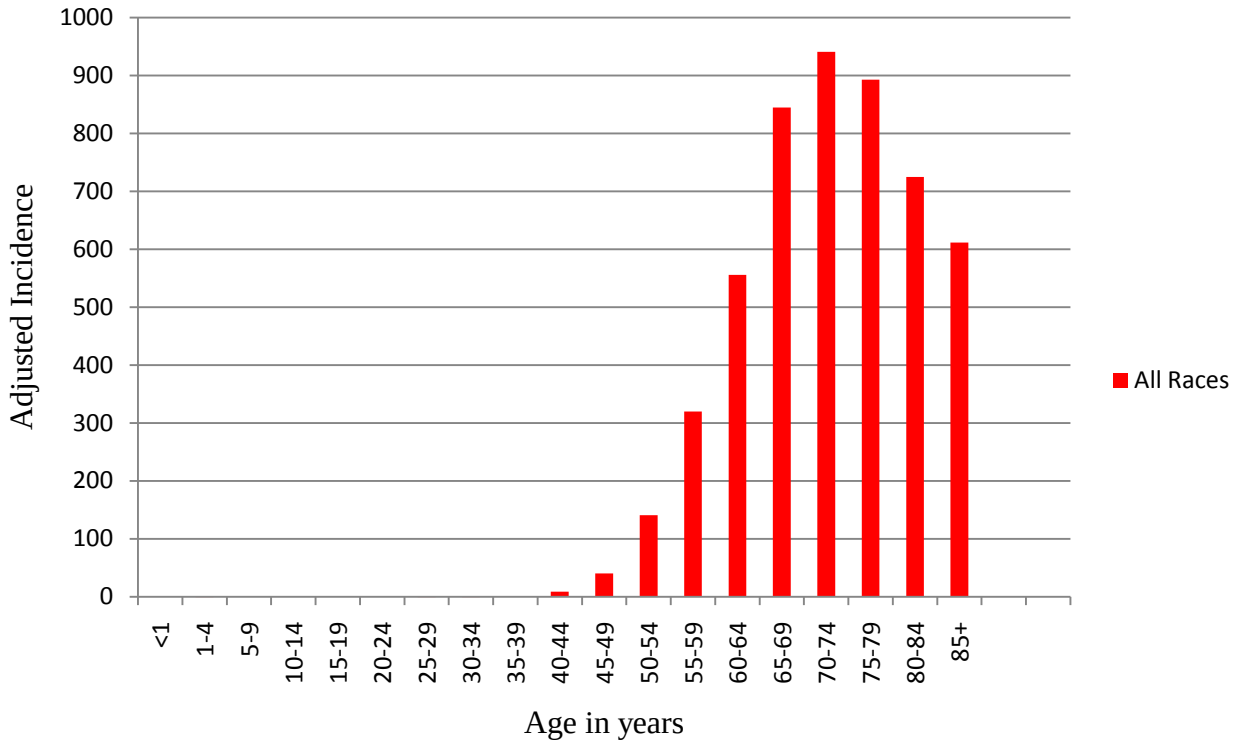
Cancer of the prostate gland is the second leading cause of cancer related deaths in men in the United States [1]. In 2009, it is estimated that nearly 193,000 men were diagnosed with cancer of the prostate gland and nearly 27,000 of them will die because of this cancer [2]. Nearly one in every six men will develop cancer of the prostate gland in his lifetime. Prostate cancer is a significant problem in the United States not only because of the numbers of men affected but also the cost of treatment of the disease is staggering.

The development of cancer of the prostate gland is influenced by a milieu of factors, including heredity, race, age, diet, physical activity, sexual factors and obesity [3]. Age is an important risk factor that affects the likelihood of contracting the disease. As a man ages, his risk of developing prostate cancer increases. The incidence of prostate cancer peaks around the age of 70 and begins to decline slightly [4]. The mean age of death of a man is approximately 76 years [4]. Figure 1-1 displays the incidence of prostate cancer per 100,000 patients broken down by age, the data for the figure was obtained from the National Cancer Institute Surveillance Epidemiology and End Results database. Another important risk factor that affects the development of prostate cancer is the race of the patient. Incidences of prostate cancer are significantly higher for an African-American than they are for a white male. African- American men have a higher incidence of prostate cancer at every age than their white counterparts [4] . The non-age adjusted incidence of prostate cancer for men in the United States is 268 out of every 100,000 men.

Figure 1-1: Incidences of prostate cancer vary with the age of the patient

The incidence of prostate cancer varies with the age of the patient. Cancer of the prostate gland is exceedingly rare in men under the age of 50. As a man continues to age, the incidence of prostate cancer begins to increase dramatically and peaks at approximately 70 years of age [4].

Figure 1-1: Incidences of prostate cancer vary with the age of the patient



Caucasian males suffer a lower than average incidence of prostate cancer at a rate of 251 men diagnosed for every 100,000. African-Americans have a much higher incidence of prostate cancer, 385 out of every 100,000 men will suffer from cancer of the prostate gland. Figure 1-2 shows the overall incidence of prostate cancer broken down by age and race [4].

The incidences of prostate cancer have experienced dramatic fluctuations in rates of disease diagnosis. The non-age adjusted rate of incidence per 100,000 persons is plotted for each year from 1975 – 2006 in Figure 1-3. From 1975 to 1994 the diagnosis of prostate cancer increased significantly. This increase in diagnosis of the disease is in part because the average life span of Americans has increased and as previously mentioned men suffer from an increased risk as they age. Other factors that may have lead to an increased rate of diagnosis are an increased awareness of the disease and an increased availability of diagnostic techniques that enable physicians to more easily detect the presence of a diseased prostate gland [5]. In 1986, the food and drug administration (FDA) approved the use of the prostate specific antigen (PSA) test to monitor the progression of cancer. In 1994, the FDA extended the utility of PSA analysis, allowing physicians to utilize the level of PSA in the blood as a tool for the diagnosis of prostate cancer. Prior to approval of the PSA screening tool, physicians were limited to diagnosing prostate disorders by digital rectal examination of the gland.

Digital examination of the prostate gland is less effective than PSA for detection of tumors [6]. In a multi-centered comparison of digital examination to serum levels of prostate specific antigen, it was shown that the serum PSA level has a 32% positive predictive value of diagnosing cancer of the prostate. While the digital rectal examination, only has a positive predictive value of 21%. If the two methods are utilized

Figure 1-2: Incidences of prostate cancer vary by age and race

Cancer of the prostate gland is more common in men around the age of 70 than at any other age. For each age group, an African-American male is much more likely to get cancer of the prostate than his white counterpart [4].

Figure 1-2: Incidences of prostate cancer vary by age and race

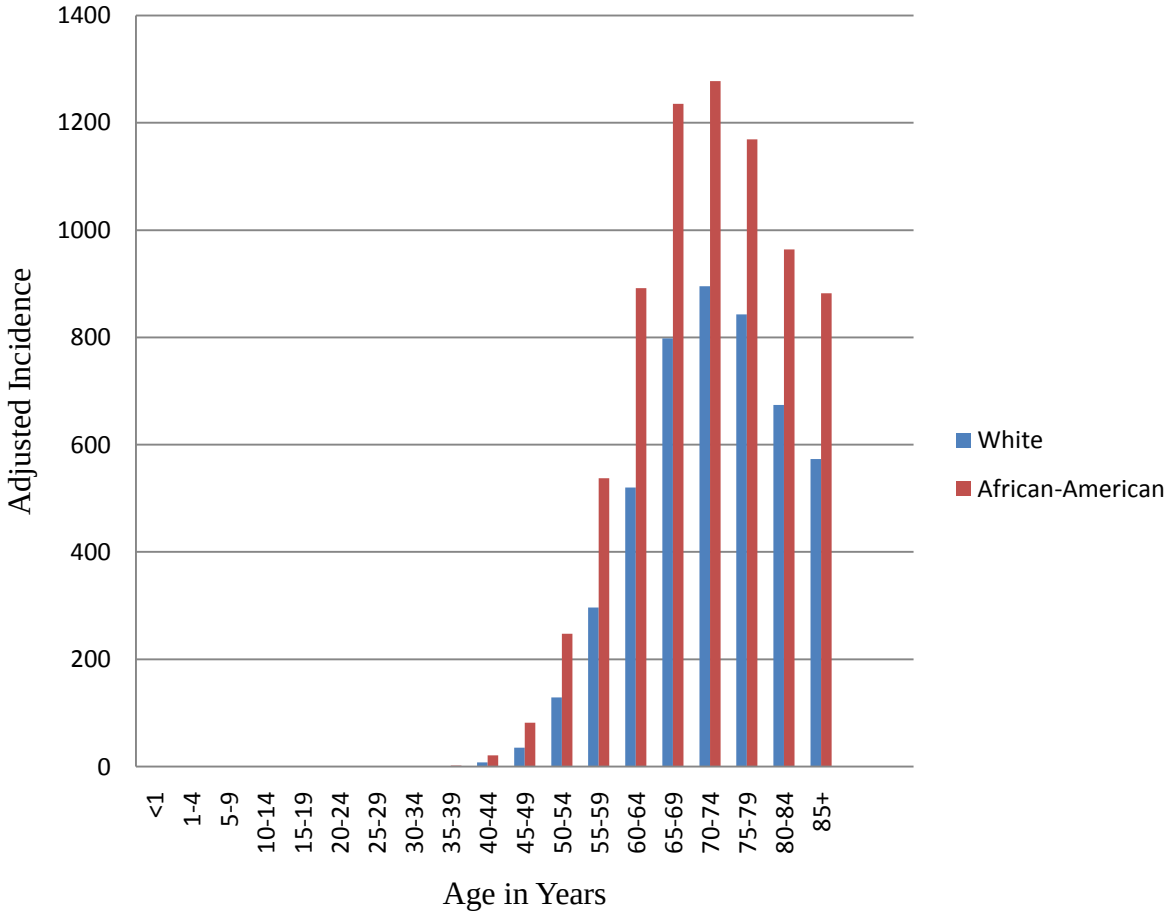
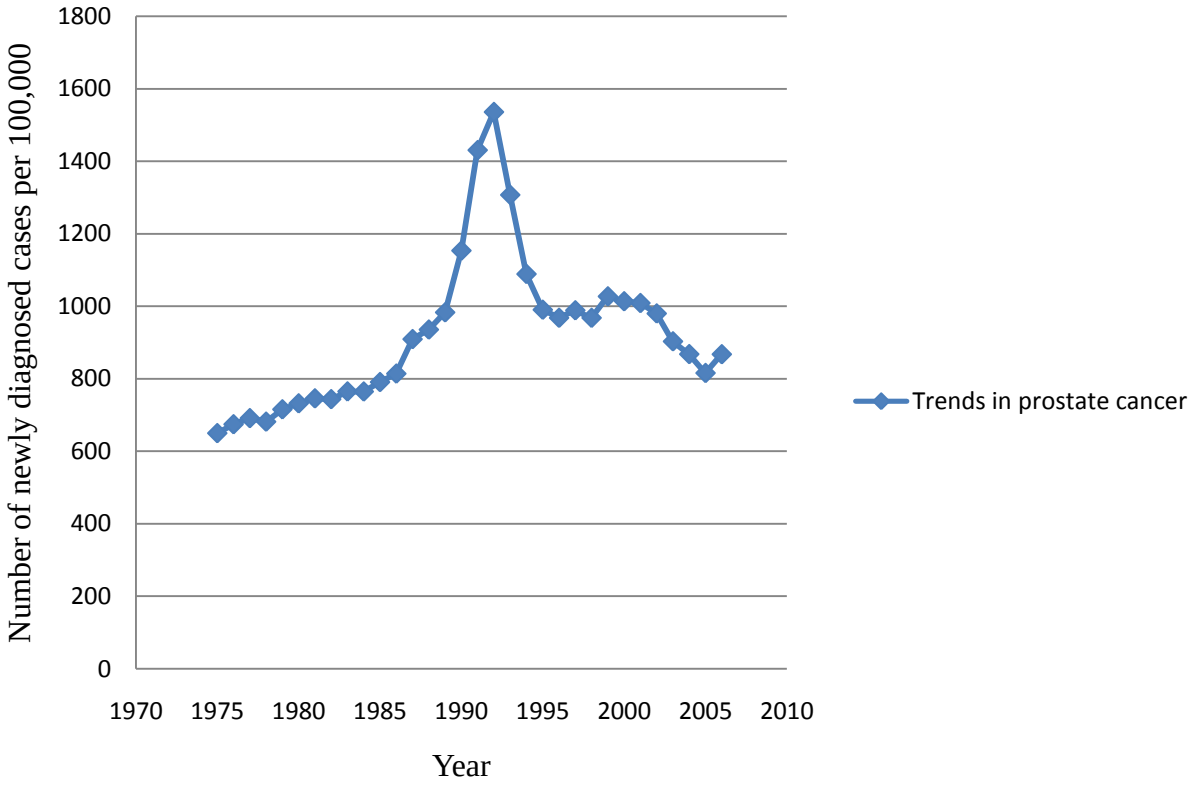


Figure 1-3: Trends in prostate cancer diagnosis

In 1975, approximately 600 of every 100,000 men without regard to their age were diagnosed with prostate cancer. In the middle part of the 1980's, the numbers of persons diagnosed with prostate cancer began to grow exponentially. It is hypothesized that the increase in diagnosis of patients with the disease is a result of an increase in the average life span, an increased awareness of the disease, and the advent of the PSA screening tool [6]. Even though it was not officially endorsed by the FDA, it is thought that many physicians were using the PSA level as a method to screen for potential tumors [5].

Figure 1-3: Trends in prostate cancer diagnosis



simultaneously, the positive predictive value increases substantially to nearly 78%.

The role of genes in prostate tumorigenesis

Tumors of the prostate gland are highly heterogeneous clinically and histologically[7]. Despite the heterogeneity of the tumors, it has been noted that several genes may play a role in the development of prostate cancer and tumorigenesis. Highly metastatic androgen independent tumors were found to exhibit point mutations in the androgen receptor approximately 50% of the time [8]. Testosterone binds to the androgen receptor and stimulates transcription of the androgen responsive genes. Genes regulated by this manner cause the cells of the prostate to grow. Androgen independence has long been suspected to play a role in oncogenesis of the prostate.

There are many other genes that are suspected to play a part in tumorigenesis. A recent study that attempted to identify a set of biomarkers to diagnose and stage tumors of the prostate found several genes to be differentially regulated in prostate cancer [9]. Table 1-1 highlights some of the most differentially regulated genes identified. Clearly, many genes are thought to play a role in the development of prostate cancer through either point mutations that inactivate the gene or through functional dysregulation.

It was recently discovered that there exist a novel class of small RNA elements that play a large role in gene regulation by blocking translation or marking the transcript for degradation [10]. These small RNAs are better known as microRNAs (miRNAs, miRs).

Table 1-1: Genes predicted to regulate tumorigenesis of the prostate

Genes that are differentially regulated between normal prostate cells and cancerous cells of the prostate gland [9]. The function of the genes was inferred by examining the information contained in the Information Hyperlinked over Proteins database [11]. The chromosomal location of each gene was examined in order to identify potential genes that lie in regions of the chromosome with a known loss of heterozygoisty involved in prostate tumorigenesis. Chromosomal locations were determined using the Entrez cytogenetic band information from the GeneCards resource [12]. Each of the genes listed in this table are suspected to play a role in the development or progression of prostate cancer.

Table 1-1: Genes predicted to regulate tumorigenesis of the prostate

Gene Symbol	Gene Name	Function	Chromosome
IGFBP-5	Insulin Like Growth Factor Binding Protein 5	Binds to insulin like growth factors and modulates cell growth; upregulated in metastatic cancer of the prostate	2q33-36[13]
FAT	FAT Tumor Suppressor Homolog	Cadherin related tumor suppressor	4q35 [13]
RAB5A	RAS related protein 5 B	Member of the RAS oncogene family; regulates vesicular trafficking	3p24-22 [13]
MTA1	Metastasis associated protein 1	Histone deacteylase inhibitor	14q32.3 [13]
MYBL2	v-myb myeloblastosis viral oncogene homolog (avian)-like 2	Transcription factor involved in cell cycle progression	20q13.1 [13]
HPN	Hepsin	Cell growth inhibitor	19q11-13.2 [13]
PIM1	Pim-1 oncogene	Proto-oncogene Serine/ threonine protein kinase	6p21.2 [13]

Discovery of microRNA and their role in development and disease

MicroRNA molecules are short, endogenous molecules that play very important roles in gene regulation by modulating protein levels in the cell [14]. It appears that microRNAs bind to complementary sequences in the target gene and repress protein translation or mark the protein for degradation. First discovered in 1993 in *C. elegans* as temporal regulators of worm development, microRNAs have been found to be ubiquitous in eukaryotes[15]. The Sanger microRNA database was created in 2002 with a list of 218 microRNAs obtained from direct submission by researchers [16]. In September 2009, version 14.0 was released with a total of 10,883 unique microRNA sequences in a variety of organisms. The Sanger microRNA repository currently lists 772 human microRNA molecules. It is hypothesized that there are many other microRNAs that have yet to be located in the human genome. miRNAs are involved in numerous cellular functions and have been shown to be involved in many critical and diverse cellular processes from cellular differentiation, viral defense, and regulation of cellular signaling networks [17]. Researchers have just begun to unravel the functions of a few microRNA molecules in everyday cellular processes and disease progression. There remain a large number of miRs that have yet to be explored.

Mature microRNAs are approximately twenty two nucleotides in length and are thought to bind to the 3' untranslated region of the messenger RNA and guide the RNA induced silencing complex to the message. After the microRNA binds to the mRNA, the mRNA is translationally repressed or degraded [16]. Elucidating the exact role of microRNA regulation and dysfunction in disease continues to be a complicated undertaking.

microRNA role in cancer progression

microRNA molecules play a role in the development of several forms of human cancer including breast, prostate, lung, thyroid, and B cell lymphoma [18]. Several microRNAs regulate critical biological processes such as cell proliferation control, cell hypertrophy, apoptosis, cell survival and insulin secretion [19]. As cancer is the end result of uncontrolled proliferation and survival of damaged cells, these biological functions have previously been shown to contribute to the development and progression of cancer.

MicroRNAs can contribute to oncogenesis by functioning as oncomiRs or tumor suppressors. miRs that regulate genes controlling cell proliferation, hypertrophy, and angiogenesis are often considered to function as tumor suppressors [20]. Loss or deregulation of tumor suppressing microRNAs imparts a growth or survival advantage to the cells, resulting in the formation of tumors. miRs that regulate apoptosis are often considered oncomiRs, increased levels of oncogenic microRNAs will impart an advantage to the cells and lead to increased tumor formation.

microRNA-17-3p functions as a tumor suppressor in the prostate

microRNA-17-3p has been shown to affect the tumorigencity of the prostate gland [21]. An *in vitro* cancer progression model system of genetically related prostate sublines showed increasing or decreasing levels of miR17-3p that negatively correlate with the oncogenic nature of the tissue [21]. The parental P69 cell line is a tumorgenic, non-metastatic cell that shows relatively high levels of human microRNA-17-3p. The highly metastatic cell line, M12 showed a two-fold decrease in the level of microRNA-17-3p. The M12 subline has been shown to contain

a loss of one copy of chromosome 19p-13 that resulted from an unequal translocation of chromosome 16:19. The F6 subline is a poorly tumorigenic, non-metastatic cell line that resulted from the micro-cell fusion techniques for restoration of the second copy of chromosome 19. The F6 subline expresses higher levels of miR17-3p than the parental P69 (Figure 1-4). Restored expression of miR-17-3p in the M12 cell line was shown *in vitro* and *in vivo* to reduce tumorigencity by at least 50% [22].These experiments clearly show that microRNA-17-3p functions as a tumor suppressor *in vitro* [21].

Clinical human samples derived from formalin-fixed, paraffin embedded samples obtained after prostatectomy confirmed that levels of microRNA-17-3p decrease as tumorigencity increases [21]. Relative levels of miR-17-3p were significantly decreased in regions of the prostate that were cancerous. Further, it has been shown that levels of miR17-3p decrease as the Gleason score of the tumor increases. Essential to understanding the role of the microRNA in the tissue is the identification of putative targets of the microRNA. Previous studies show that miR17-3p regulates expression of vimentin [22].

Role of intermediate filament proteins in cancer

Intermediate filaments are fibrous proteins found within the cytoplasm and nucleoplasm of most eukaryotic cells [23]. Some intermediate filaments like vimentin tend to be peri-nuclear localized and form a cage around nucleus extending to the surface of the cell. It has long been known that intermediate filaments are essential for the internal integrity of the cell and the shape of the cell. However, it has become clear in the past few years, that intermediate filament proteins are dynamic molecules involved in many regulatory functions.

Figure 1-4: Genetically related prostate cancer progression cell lines

Unique genetically related cell line, derived from injection of prostate cancer cells into nude, athymic mice [24]. The parental cell line (P69) is a slightly tumorigenic, non-metastatic cancer cell line. After several rounds of injection, the cells became highly tumorigenic and metastatic (M12). The subline was noted to exhibit an unequal translocation of chromosome 16:19. Microcell fusion techniques were used to inject the missing region of chromosome 19 into the M12 subline [25]. After insertion of the missing region of chromosome 19, the subline became less tumorigenic and non-metastatic (F6).

Figure 1-4: Genetically related prostate cancer progression cell lines

Name of cell subline	Description	Relative levels of vimentin	Relative levels of miR-17-3p
P69	Parental cell line, tumorigenic and non-metastatic	0.09	0.0032
M12	Highly tumorigenic and metastatic subline; Ch 16:19 translocation	0.24	0.0018
F6	Restored chromosome 19; non-tumorigenic	0.01	0.0038

Vimentin is an intermediate filament protein that is produced by the *Vimentin* gene on chromosome 10p13 [12]. The protein is 466 aminoacids in length and has a mass of 53.6 kDa. Vimentin has been shown to be differentially regulated in prostate cancer cell lines. Previous studies have demonstrated that tumors containing higher levels of vimentin are more motile and invasive [26]. The highly metastatic M12 cell subline has significantly higher levels of vimentin compared to the less tumorigenic P69 and F6 cell lines [22]. Likewise, vimentin expression was significantly increased in highly invasive, androgen insensitive cell lines (LnCap CL1) [26].

Subcutaneous injection of tumor cells into nude, athymic, male mice revealed that the M12 subline is highly tumorigenic *in vivo* [22]. The mice injected with the M12 subline all formed tumors within nine to fifteen days. The F6 subline injected mice failed to grow a tumor or grew very small tumors 120 days post-injection. As the levels of vimentin vary among the cell lines, the effect of vimentin on tumor growth was investigated by injecting mice with M12s along with a vimentin small hairpin RNA (shRNA). The mice injected with the M12 variant expressing a vimentin shRNA (M12 +siVim). In the mice injected with the M12 + siVim variant the size of the tumors was greatly reduced compared to the M12 subline alone. To investigate the role of human microRNA17-3p on tumor formation in nude, athymic mice, the animals were injected with an M12 variant subline containing a microRNA-17-3p over expression vector [21]. Tumor formation was reduced but not to the same degree as the mice containing tumors with the shRNA for vimentin. It seems that miR-17-3p may regulate some protein(s) that infers a slight advantage to the cells. In order to further understand the role of miR-17-3p in the prostate, one must identify these other putative targets of the miR, in addition to vimentin.

Biogenesis of microRNA molecules

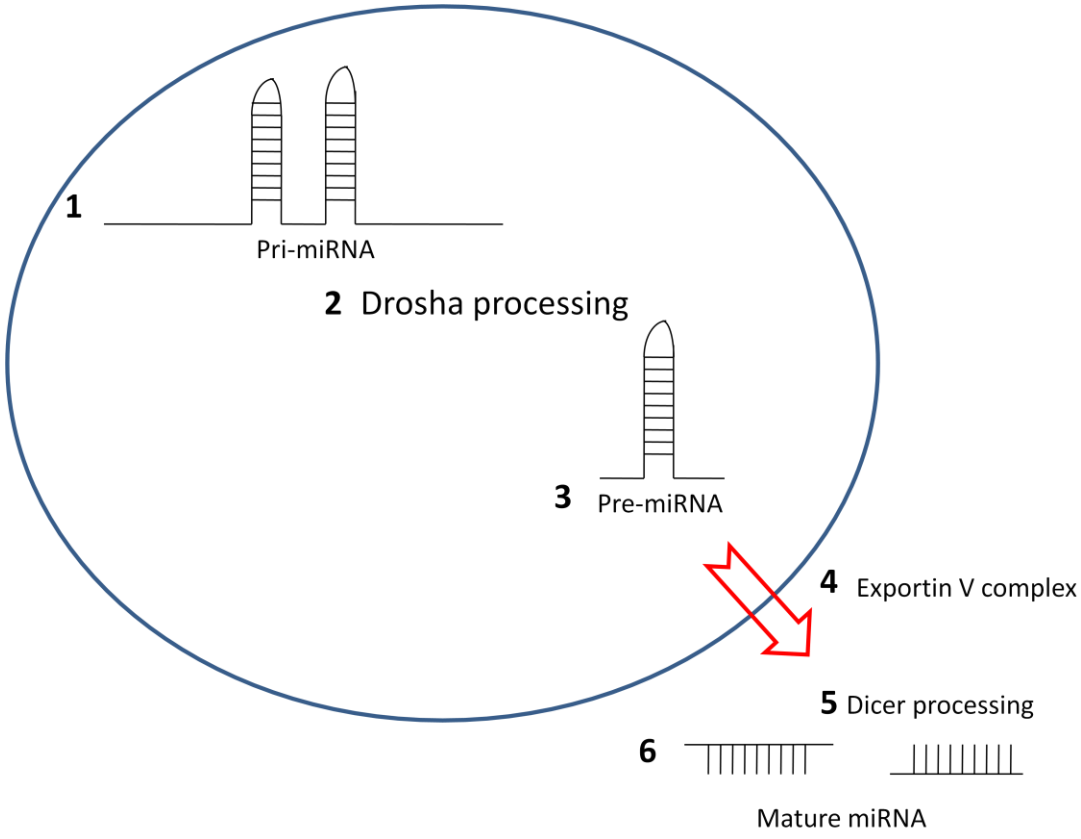
Essential to understanding the role miRNAs play in cellular processes is the identification of their putative targets. The past decade has seen a large increase in the numbers of computational methods for the identification of miR/ gene targets. The methods employed by the various computational tools are diverse. They range from simple string based methods to more complicated markov models. Before understanding the techniques of microRNA target prediction, one must understand the biogenesis of microRNA and the types of interactions between miRNAs and their target genes.

miRNA genes can be part of a polycistronic transcript of 2- 7 microRNA genes under the control of a common transcriptional regulator or can be excised from introns of protein coding genes [27]. MicroRNAs in animals are created using a two step process [28]. Figure 1-5 illustrates the step wise biogenesis of the microRNA molecule. The first step involves transcription of a several hundered nucleotide in length transcript that is called the pri-miRNA by RNA pol II or RNA pol III. The pri-miRNA is processed into a smaller structure of approximately 70 nucleotides (pre-miRNA) in length while in the nucleus. The processing is accomplished by a complex of several proteins, the most important being Drosha [29]. The pre-miRNA has a characteristic stem loop shape and is exported from the nucleus by a complex of Exportin 5 of Ran-GTP [30]. Once in the cytoplasm, the pre-miRNA undergoes the second phase of processing, where it is processed by the enzyme Dicer into its mature form that is 22 nucleotides in length [31].

Figure 1-5: Biogenesis of microRNA

1. MicroRNA molecules are transcribed from introns of protein coding genes or are a part of a polycistronic transcript by RNA polymerase II or RNA polymerase III. The pri-miRNA is several hundred nucleotides in length and contains regions of stems and loops. [27]
2. A complex of enzymes that includes Drosha, a Rnase II enzyme, cleaves the pri-miRNA into the shorter microRNA molecule (pre-miRNA) [27, 32].
3. The pre-miRNA has a characteristic stem loop shape and is approximately 70 nucleotides in length [32].
4. The pre-miRNA is exported from the nucleus to the cytoplasm by the Exportin V complex [30].
5. In the cytoplasm, the enzyme Dicer recognizes the characteristic stem loop shape of the pre-miRNA and cleaves into the functional mature microRNA molecule [31].
6. Mature microRNAs can be generated from either the 5' or the 3' end of the stem loop of the pre-miRNA [29]. The mature micrRNA is typically 22-25 nucleotides in length and binds to the 3' untranslated region of the mRNA transcript.

Figure 1-5: Biogenesis of microRNA



Computational methods of identification of putative targets

MicroRNAs interact with their targets through sequence complementarity. In some cases, seven complementary bases on the 5' end of the microRNA are adequate to post-transcriptionally regulate protein levels [33]. Bases 2-7 of the microRNA seem to be critical in order for the microRNA to be able to bind to its target. These bases are commonly referred to as the seed region of the miRNA. Many miRNA/ target interactions exhibit 3' compensatory loop interactions that further stabilize the structure of the molecule. Due to the relatively short size of microRNA molecules and the small number of complementary bases, computational prediction of potential targets of miRNAs is very difficult [34]. Other features that complicate attempts at computational prediction of interactions are the presence of regions of mismatches between the microRNA and the target 3' UTR, G:U wobbles and sequence bulges [35]. Table 1-2 lists the details of some of the most commonly utilized tools for prediction of microRNA targets.

The miRanda algorithm

One of the earliest attempts to identify potential putative targets of microRNAs was the development of the miRanda algorithm [36]. The algorithm employed by miRanda utilizes a two step process. First, it inputs the sequence of the microRNA and compares it to the 3' untranslated region of all genes in the input file. Seed region base complementarity is given a higher reward than complementarity of bases on the 3' end of the microRNA molecule. The reason for this reward is because it is generally thought that the seed region interactions are critical to the ability of a miRNA to repress protein translation. The program evaluates each potential match based on base sequence complementarity. Complementary sequences that exceed a pre-defined

Table 1-2: Common MicroRNA/ Target Prediction Tools

Computational methods to identify potential putative targets of microRNA are essential to elucidation of their role in the biological realm. We list some of the most commonly cited computational tools in existence, the world wide web address of each tool and the number of times each prediction tool has been cited in microRNA related literature as of September 2009. Often the number of citations is used to indicate popularity of a program. Utilizing this metric, it appears that PicTar and TargetScan are the most commonly utilized tools for microRNA target identification.

Table 1-2: Common MicroRNA/ Target Prediction Tools

Tool	Web Location	Type of Tool	Times Cited	Reference
Diana MicroT 3.0	http://diana.cslab.ece.ntua.gr/microT/	Precompiled List	Not Available	[18]
Miranda	http://www.microrna.org/microrna/home.do	Open source and precompiled list	30	[36]
PicTar	http://pictar.med-berlin.de/	Precompiled List	595	[37]
TargetScan	http://www.targetscan.org/	Open source and precompiled list	1081	[38]
RNA 22	http://cbcsrv.watson.ibm.com/rna22.html	Precompiled List and webserver	114	[39]

.

threshold set by the user will enter the second phase of evaluation. Potential matches will be evaluated for thermodynamic stability using the RNAlib module of the Vienna RNA package [40].

TargetScan

TargetScan employs an algorithm that scans the 3' UTR of possible targets for the presence of one of three types of canonical seed region interactions [38]. The most favorable and highly scored interaction results from complementarity of positions 2-8 of the microRNA and the presence of an adenine nucleotide at the far 3' end of the potential binding site in the UTR. The second most favorable interactions come from binding of bases 2-7 of the microRNA, along with an adenine residue at the end of the bound region of the target or binding of bases 2-8 of the microRNA to the target. The least favorable target considered to be possible utilizing the TargetScan algorithm is complementarity of bases 2-7 of the microRNA and target UTR.

PicTar

PicTar is a computational tool for the prediction of microRNA targets that not only employs seed region interactions, but also ensures evolutionary conservation and secondary structure stability [37]. PicTar accepts as input two files; the first contains a multiple sequence alignment of the 3' UTR of related species and a second file containing the microRNA sequences of the researcher. The algorithm employed by PicTar searches the multiple sequence alignment for sequences that would adhere to the seed region by strictly following the Watson-

Crick base pairing rules. The algorithm allows the seven bases of the seed region to begin at either position one or two of the microRNA. Potential microRNA binding sites are filtered by free energy of heteroduplex formation. Binding sites that have a free energy lower than the desired threshold are retained and further analyzed. If the potential binding site exists in a region that is completely conserved across several species, the binding site is considered to be valid and a hidden markov model maximum likelihood fit score is calculated. Potential binding sites are ranked based on their score. Higher scores represent binding sites that are more likely.

Diana MicroT 3.0

Diana microT 3.0 utilizes sequence similarity, free energy of heteroduplex formation and to a lesser degree multiple species conservation to identify potential binding sites [18]. The algorithm employed by Diana, pulls out the first nine nucleotides (driver sequence) and uses a sliding window approach to scan the 3' UTR of the gene for sequences that are complementary to the driver sequence in at least six consecutive nucleotides. The program allows a single G:U wobble, as long as there are at least six Watson-Crick base pairs in the alignment.

Potential binding sites with less than seven Watson- Crick pairs are further evaluated using the free energy value of heteroduplex formation using the RNA hybrid algorithm [41]. The first step of filtration is accomplished by passing the actual microRNA sequence and the potential binding site to the RNA hybrid program. Following the calculation of the actual binding energy, a hypothetical binding energy is calculated by estimating the binding energy of the perfectly complemented sequence to the microRNA. If the ratio of theoretical to hypothetical binding energies is greater than 0.74, the sequence is identified as a potential "miRNA

recognition element (MRE)" or a predicted binding site [18].

Sequences of multiple species are examined to determine if the potential MRE is present in other species to identify sequences that are conserved in multiple species [18]. It is hypothesized that sequences that have been conserved though out millions of years of evolution possess functional significance. Species that are examined include humans, rats, mice, dogs and zebrafish. A conservation score is generated that is equal to the number of species that contain the individual MRE. The conservation score and the binding score are combined into a microRNA target gene score (miTG). Potential binding sites are ranked by their miTG scores.

RNA22

RNA22 is a pattern based method that relies on information inherent in the nucleotide sequences of the microRNA molecule [39]. RNA22 uses the Teiresias algorithm to identify patterns in the sequences of microRNA molecules from reference species in the RFAM database [42, 43]. Following pattern identification, the UTRs of the genes of interest are scanned for the presence of "target islands" that possess at least one of the patterns identified in the pattern matching step of the algorithm. Target islands are matched to candidate microRNA molecules based on sequence complementarity. A score, based on the number of Watson-Crick base pairs in the heteroduplex, is generated. Potential binding sites that exceed a given threshold are returned to the user.

Project Objectives

The computational identification of putative targets of microRNA is critical to understanding the role a given miR plays in development and disease. Presumably, programs that

include multiple features of miR/ target interactions will be more accurate than programs that include fewer features. In practical terms, accuracy is defined as the ability of a method to identify true interactions and reduce the number of false predictions. Researchers desire to identify the greatest number of true interactions at the lowest cost of investigation in terms of fiscal cost and manpower. With so many computational methods in existence, researchers must understand the features, advantages, and benefits of each of the programs.

In order to fully understand the mechanism of tumorigenesis, this project seeks to identify putative targets of miR-17-3p using a combination of computational and traditional wet-lab techniques. Prior to the utilization of previously published microRNA prediction tools, all tools were evaluated using a standardized set of proven microRNA and gene targets to measure the accuracy of the programs. As part of this project, we designed and implemented a comprehensive microRNA annotation and prediction interface (MAPI) that increased the accuracy of current programs. MAPI was utilized to identify potential targets of miR-17-3p that are expressed in the prostate gland and involved in tumorigenesis. Utilizing wet-lab techniques, the targets were evaluated in a cancer cell progression model using western blot analysis of target protein levels.

Specific Project Aims

- Evaluate current computational microRNA prediction tools to determine which program or combination of programs offer users the best balance of sensitivity and specificity.
- Design and implement a comprehensive computational interface that increases the effectiveness of previously published programs by filtering irrelevant targets
    - Include transcriptional profiles for various tissues and disease states
- Use the computational interface to identify potential targets of human microRNA-17-3p that potentially impart an oncogenic advantage to the cells when overexpressed.
- Utilizing a cancer cell progression model, verify that the levels of the predicted target of microRNA-17-3p are more abundant in the highly tumorigenic and metastatic M12 cell line.

Chapter 2

Evaluation of computational methods of microRNA target identification

Methodology to compare microRNA prediction programs

The past decade has seen a large increase in the attempt to determine the exact role microRNAs play in the maintenance of health and the development of disease. To understand the role of a given microRNA in human disease, a researcher must first identify its putatitive targets. Many computational tools/ programs have been developed to assist researchers in the prediction of microRNA target gene interactions. The methods employed by the various computational programs vary greatly. This project seeks to evaluate the sensitivity of each method along with determining the total number of predictions. This work compares the predicted targets of Diana MicroT, Miranda, Pictar, TargetScan, and RNA22 utilizing a set of microRNA molecules that are shared by all methods [18, 37-39, 44]. This work only evaluates the performance of the various programs to predict human microRNA gene interactions as most researchers are interested in determining the role microRNA play in the development of human disease.

A previous study conducted by Sethupathy, Megraw, and Hatzigeorgiou in 2006 evaluated the sensitivity of various target identification programs using a benchmarking dataset of targets from the Tarbase database for experimentally proven microRNA gene target pairs [45, 46]. At the time of that publication, there were only 84 such targets for 32 microRNAs that had been proven. In the past three years, both the number of putative targets for microRNAs and the number of prediction programs have increased dramatically. It is reasonable to suspect that with the increase of available data and more advanced methods that the former conclusions are no longer valid.

Researchers found that the single best tool in existence in 2006 was Miranda and that the best sensitivity could be achieved by overlapping the predictions from every program available

[45]. However, biologists cannot search through the number of predictions generated by overlapping the predicted targets. Scientists often attempt to enrich the number of true interactions by considering only microRNA targets that have been predicted by more than one method. The previous study recommended that in order to balance the needs of researchers to maintain a high level of sensitivity and allow researchers to search through a lower number of potential targets, only targets predicted by PicTar and TargetScanS be considered [37, 38].

Compilation of Predictions

Predictions for DIANA-MicroT 3.0 were downloaded in a pre-compiled tab delimited text file from their webserver, which can be found at the URL listed in Table 1. Human microRNAs that target a unique gene with a miTG score of at least 1.0 were included in our set of predicted targets. Multiple interactions per gene, if possible, were collapsed into a single predicted pairing. PicTar predictions were compiled using a PERL script that downloaded all predicted targets for each human microRNA that was available on the PicTar server as of July 5, 2009. TargetScan 5.1 human data was downloaded from the TargetScan server using the predicted conserved targets file. Precompiled predictions were downloaded from Miranda using the web address specified in table 1 and downloading the human miRNA target site predictions file. RNA 22 predictions were assembled using the precompiled data file for only the 3'-UTR region.

Development of a standardized microRNA/ target comparison set

In order to measure the sensitivity of each computational program, we assembled a set of microRNA/ target interactions that have been experimentally proven. Tarbase and miRecords are web based resources where researchers report microRNA/ gene pairs that have been experimentally validated [46, 47]. By combining the two datasets and eliminating multiple entries, we developed a comprehensive list of proven human microRNA/ gene interactions. MicroRNAs have been shown to bind to more than one site on a given target gene, in this study we eliminated such multiple binding sites and considered the pairing as a single predicted interaction. We compiled a list of 826 experimentally verified microRNA targets.

Sensitivity Analysis

As the number of microRNAs vary slightly between tools, it was necessary to create a standard set of microRNA molecules that could be used to accurately compare each tool. We intersected the microRNA identifiers from each tool and found that only 139 microRNAs were shared among all of the tools that we compared in this study. Of these 139 common microRNAs, we found that there were 451 unique interactions. These interactions were the only predictions that we utilized in the sensitivity analysis.

Each prediction method previously described was compared to the standardized comparison target set in order to determine how well each tool performed. Of particular interest to microRNA researchers is the sensitivity of each tool. Figure 2-1 shows the formula for derivation of sensitivity. Sensitivity evaluates the proportion of experimentally proven

interactions that are correctly identified by the evaluated program. Successful tools will ideally be able to correctly identify a majority of the experimentally verified microRNA targets.

Following individual program analysis, we evaluated various combinations of tools to determine if there exists an ideal combination of computational tools that can identify most microRNA targets. In order to accomplish this objective, each set of microRNA/ target predictions was pair wise intersected with all other tools in an effort to determine the sensitivities of each possible combination of tools. Likewise, sets of predictions for each tool were overlapped (unioned) in pair wise fashion with every other tool in an attempt to increase the sensitivity. We illustrate the concept of mathematical union and intersection in Figure 2-2.

Many researchers base their decision on which microRNA tool to use based solely on the sensitivity of the various tools. However, one could theoretically predict every gene to be a target of each microRNA and achieve a sensitivity of 100%. Specificity measures the proportion of incorrect or negative interactions that are correctly identified. It is difficult to measure the specificity of a microRNA target prediction tool, as there is not a comprehensive database that tracks negative experimental results. Therefore, as a surrogate for specificity, in this study we compared the performance of each tool by calculating the total number of predictions. The numbers of predictions are only the predicted pairings that originate from our set of 139 common microRNA molecules.

Figure 2-1: Sensitivity Analysis

Details the analysis process utilized to measure sensitivity of each microRNA prediction tool and combination of prediction tools.

Figure 2-1: Sensitivity Analysis

Sensitivity is defined as:

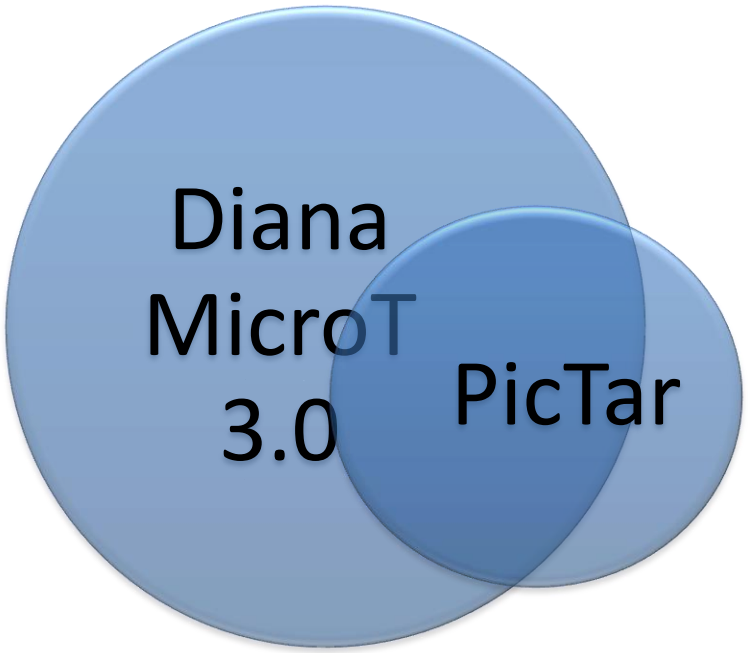
$$\frac{\text{Number of true positive interactions}}{(\text{Number of true positive interactions} + \text{number of false negative interactions})}$$

True Positive Interaction	Interaction predicted by the tool or combination of tools and is found to be in our standardized list of experimentally proven targets
False Negative Interaction	A member of the standardized list of experimentally proven interactions that is not correctly predicted by the tool or combination of tools.

Figure 2-2: Venn diagram of the union and intersection of the PicTar and Diana MicroT 3.0 datasets

Targets were predicted for the 139 microRNAs common to all tools. Diana MicroT 3.0 predicts approximately 373,000 total interactions among our standardized list of microRNAs. PicTar predicts slightly less than 49,000 interactions among the 139 microRNAs in common to all prediction programs compared. If a researcher only considers the microRNA targets predicted by both programs (targets in the intersection of the two tools), the number of potential targets is reduced to approximately 29,000 possible targets. Sensitivity is increased by taking the union of the two sets of predicted targets. The union of the datasets includes any targeted predicted by either Diana MicroT3.0 or PicTar. The union of the two datasets predicts nearly 383,000 possible targets.

Figure 2-2: Venn diagram of the union and intersection of the PicTar and Diana MicroT 3.0 datasets



Findings

There exist many computational methods for prediction of microRNA/ gene interactions in humans. In this study, we compared five commonly used microRNA prediction tools. A list of tools compared in this study, along with the web location, method of data acquisition, and the number of citations referencing that tool are listed in Table 1-1.

In Table 2-1, we show the sensitivity and the number of total predictions generated using our list of common microRNAs for each tool and combination of tools. This analysis found that the single best performing microRNA/ gene prediction tool is DIANA-MicroT 3.0 [18]. Although this program has been around for several years, version 3.0 was first described in July 2009 and was included in this analysis because of their self reported 66% precision rate. Figure 2-3 plots the calculated sensitivity versus the total number of predictions for each of the tools. The algorithm utilized by DIANA-MicroT 3.0 considers several types of seed region binding, cross species conservation, and thermodynamic stability of seed region interactions. Potential binding sites are given a microRNA target gene score (miTG) which reflects the relative strength of the prediction. Our analysis found that DIANA-MicroT 3.0 was able to achieve a sensitivity rating of nearly 46%. However, to accomplish this level of sensitivity, the program predicted over 370,000 microRNA/ gene pairings. Users can decrease the total number of predictions by increasing the minimum miTG score.

Table 2-1: Sensitivity Analysis of each Prediction Tool and Combination of Prediction Tools

Comparison of the sensitivity and total number of predictions for each microRNA target prediction tool and combinations of the intersections and unions of various tools. Using this analysis it is clear that to achieve the best sensitivity, a researcher needs to combine all of the predictions of as many tools as possible. However, in order to achieve that level of precision, a researcher will need to look through nearly 3800 predictions per microRNA. The single best tool is Diana MicroT 3.0 which achieves a sensitivity of nearly 46% while predicting less than 2800 possible interactions per microRNA [18, 37-39, 44]. Programs that achieve the highest sensitivity are able to more correctly identify true interactions.

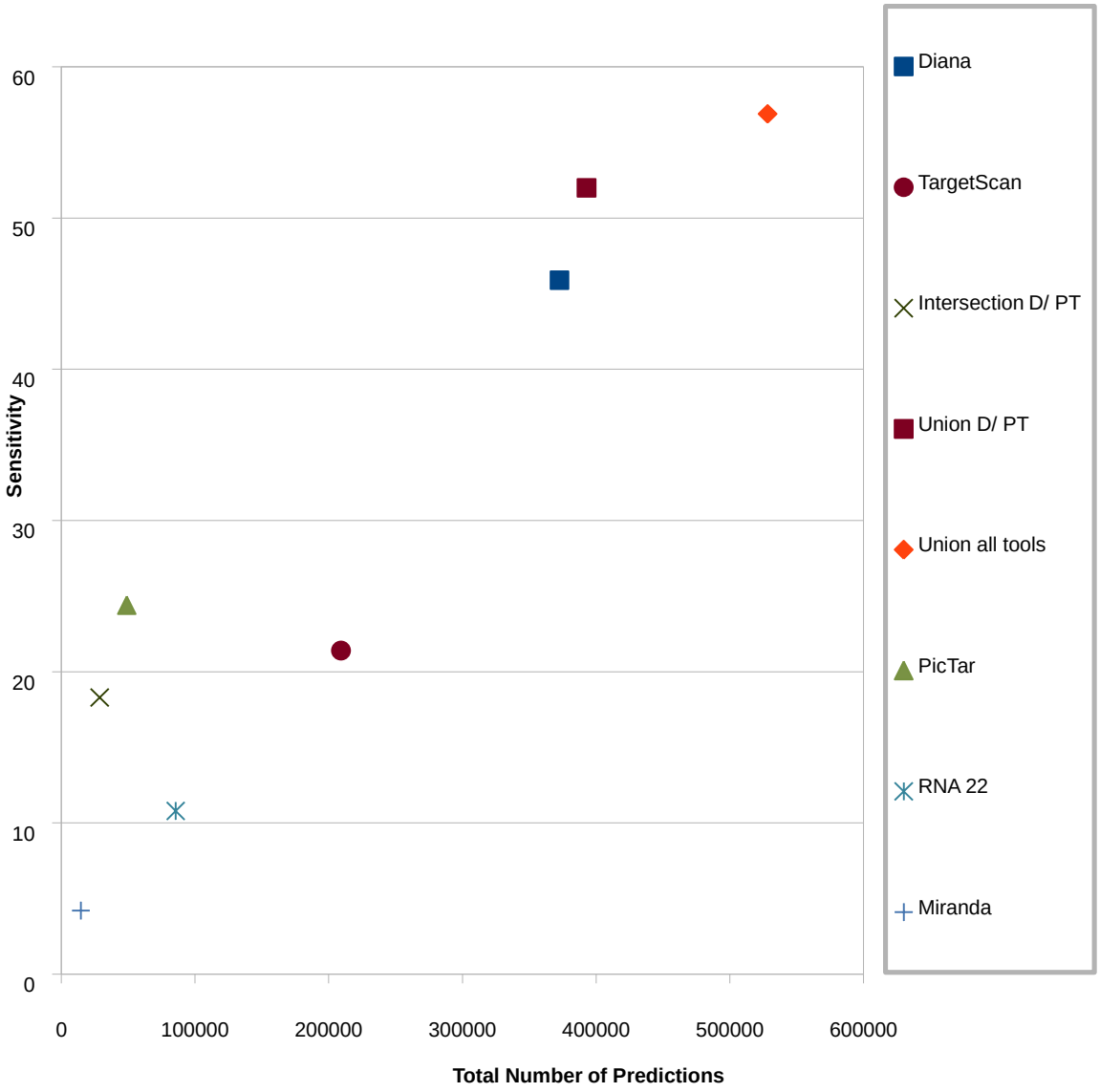
Table 2-1: Sensitivity Analysis of each Prediction Tool and Combination of Prediction Tools

Tool	Sensitivity	Total number of predictions
Diana	45.9	372530
Miranda	4.2	14612
PicTar	24.4	48948
RNA22	10.8	85481
Target Scan	21.4	209151
Intersections of tools		
Target Scan/ Miranda Intersection	2.9	4327
Diana/ PicTar Intersection	18.3	28617
Miranda/ TargetScan Intersection	2.2	3283
Diana/ Miranda Intersection	8.3	30080
Miranda/ PicTar Intersection	1.7	843
Diana/ TargetScan Intersection	17.1	131114
RNA22 / TargetScan Intersection	2.5	13659
PicTar/ TargetScan Intersection	9.7	14068
Miranda/ RNA22 Intersection	1	1089
PicTar/ RNA22 Intersection	4.7	4800
Unions of tools		
Diana/ Miranda Union	47.3	382815
Diana/ PicTar Union	52	392861
Miranda/ TargetScan Union	23.4	220480
Diana/ RNA22 Union	48.5	427931
Miranda/ PicTar Union	26.9	62717
Diana/ TargetScan Union	50.2	450567
RNA22/ TargetScan Union	29.7	280973
PictTar/ TargetScan Union	36.1	244031
Miranda/ RNA22 Union	14.1	99004
PicTar/ RNA22 Union	30.5	129629
Union of all tools	56.9	528237

Figure 2-3: Scatterplot Comparison of MicroRNA Prediction Tools

The sensitivity versus the total number of predictions for individual tools and the combinations of Pictar and Diana MicroT 3.0 was plotted [14, 37-39, 47]. Our analysis proves that the best method to identify the greatest number of experimentally proven targets is to consider any target predicted by any microRNA target prediction tool. However, using this approach researchers will have to search through an average of 3600 targets per microRNA. This is a daunting task that can only be accomplished by a large research group with abundant financial resources.

Figure 2-3: Scatterplot Comparison of MicroRNA Prediction Tools



The sensitivity of predictions of DIANA-MicroT 3.0 can be increased by adding the predictions of any other tool to the predictions made by DIANA. The union of PicTar and DIANA together are able to achieve over 50% sensitivity [18, 37]. Due to significant overlap of the predicted targets in these two datasets, this increased sensitivity can be accomplished with only a five percent increase in the total number of predictions. The best sensitivity can be achieved by compiling all of the predicted targets of each of the target prediction programs into a single list. This results in a near 57% sensitivity value, but the total number of predictions would be over a half million or nearly 3600 targets per microRNA molecule. It does not make biological sense that a single microRNA would target over one tenth of the human genome. Previous groups have suggested that it is more biologically plausible to suspect that each microRNA may target nearly 500 genes [12].

PicTar searches a multiple sequence alignment of RNA sequences for perfect seed region binding of the first seven nucleotides beginning at position 1 or 2 of the 5' end of the mature microRNA sequence [37]. The PicTar algorithm filters out potential targets that do not meet a minimal free energy cutoff using the RNA Hybrid program [48]. In addition to seed region binding and thermodynamic stability, PicTar filters out potential false positive predictions by ensuring that the UTR sequence is conserved across multiple vertebrate species. According to our analysis, PicTar is able to achieve a sensitivity score of nearly 25%. The number of potential targets predicted by PicTar is far less than the number of targets predicted by other methods. Presumably, this would make PicTar one of the more specific programs available for use by microRNA researchers.

As previously mentioned, biomedical researchers often intersect the predicted targets of multiple programs in an attempt to enrich the number of true interactions in their data set.

Recently, the validity of this approach was questioned [49]. Our analysis finds that the sensitivity of the methods is greatly decreased by intersecting combinations of predicted targets. Diana MicroT and Pictar were able to achieve 52% sensitivity when the sets of data were overlapped and unioned together. When the list of potential targets was intersected, and only targets predicted by both methods were considered, the sensitivity was reduced to 18%.

This analysis demonstrates the complexity associated with predicting microRNA targets. Researchers seeking to study the effects of a given microRNA on development or disease face a difficult task. Currently, the single best tool can only identify approximately half of the experimentally proven interactions. This finding highlights the necessity to develop newer, more accurate methods of target identification or develop methods that reduce the number of predictions associated with overlapping multiple tools.

Chapter 3

Development of the MicroRNA Annotation and Prediction Interface (MAPI)

Features of MAPI

Cancer is a highly, heterozygous disease resulting from a combination of genetic and epigenetic factors. Even though cancers are highly individual, it has been known for quite some time that each cancer has a unique molecular signature or molecular portrait [50]. This portrait of each cancer profiles the highly variable nature of gene expression as measured by the DNA microarray. Recent work has shown that expression profiles of microRNAs are also highly variable in various forms of human cancer [51]. miRNA expression profiles can be used to detect developmental lineages and differentiate between various stages of the disease. As microRNAs regulate gene translation, there exist tissue specific differences among genetic profiles in various tissues/ diseases and there exist tissue specific differences among microRNA expression levels. Thus, it is reasonable to assume that one could integrate these sources of information and yield a more accurate computational method for the prediction of targets of microRNA.

The microRNA annotation and prediction interface (MAPI) is a comprehensive tool that integrates multiple information sources into a single easy to use interface. MAPI was built using the MySQL open source database, PERL, and an html front end user interface and provides a centralized resource that includes several of the most highly referenced computational prediction programs. The database allows end users the flexibility to choose from any of the computational tools included in the program or choose from one of several combinations of programs that offer the user the highest sensitivity and specificity.

The greatest strength of MAPI is the ability to filter predicted targets based on numerous biological parameters, allowing users to retain a high level of sensitivity and decrease the overall number of predictions per microRNA. MAPI allows users to select a tissue of interest and search

for targets of a specific microRNA or search for a microRNA(s) that target a gene of interest. The program searches for microRNAs and gene targets that are co-expressed in the tissue of interest.

MicroRNA annotation

All microRNA sequences were downloaded from miRBase. miRBase is a searchable, annotated database of microRNA resources [16]. miRBase version 14.0 was used for compilation of microRNA attributes. Data was downloaded using the FTP site of the resource at ftp://mirbase.org/pub/mirbase/CURRENT/. The files downloaded included the file containing the precursor microRNA ID and sequence named hairpin.fa.gz, the file containing the mature microRNA ID and sequence named mature.fa.gz, and the file that contained the alternative microRNA from the 3' end of the precursor named maturestar.fa.gz. All files were decompressed using the unix operating system and a PERL script was written to extract the name and sequence of each human microRNA . Hairpin microRNA molecules were mapped to a chromosomal location using the hsa.gff file found in the genomes directory of the above ftp site. Each microRNA precursor was linked to the appropriate mature microRNA molecule(s). In some cases, a single precursor can give rise to two mature microRNAs. All files were assembled into database tables and uploaded to the MAPI interface (Table 3-1 and 3-2).

Table 3-1: Description of precursor microRNA table

The pre-miRNA table contains a list of all human precursor microRNA molecules. Information was obtained from the Sanger microRNA repository. Each pre-miRNA is annotated to include the chromosomal location of the precursor molecule, sequence and direction of the molecule in reference to the chromosome [16].

Table 3-1: Description of precursor microRNA table

Precursor microRNA Table
microRNA precursor ID
microRNA accession ID
Sequence
Chromosome
Chromosomal Start
Chromosomal End
Length of pre-miRNA
Chromosomal direction

Table 3-2: Description of the Mature microRNA table

The mature microRNA table contains a list of all human mature microRNA molecules, their corresponding sequence, the identifier of the associated the pre-miRNA, and the chromosome that gives rise to the mature miRNA.

Table 3-2: Description of the Mature microRNA Table

Mature microRNA Table
microRNA mature ID
microRNA precursor ID mapped to mature ID
Sequence
Chromosome

Table 3-3: Predicted targets of microRNA molecules

Each predicted interaction of a microRNA and gene are listed in the predicted targets table. The computational tool and score generated by that method are listed in the interaction table.

Table 3-3: Predicted targets of microRNA molecules

Predicted Targets of microRNA
microRNA mature ID
Target gene
Interaction score
Prediction tool

Assembly of predicted targets

Potential targets were assembled for Pictar, TargetScan, RNA22, Miranda, and Diana MicroT 3.0 [18, 36-38, 44]. The methodology described in Chapter 2 was used to compile all of the predicted targets for each prediction tool. When applicable, multiple interactions of a microRNA to a single gene were combined into a single interaction. A singleMySQL table was created from the compilation of all predicted interactions (Table 3-3).

Compilation of oncogenic genes

It is well known that aberrant expression of certain genes imparts an oncogenic nature to the cells/tissues. Through the efforts of the National Cancer Institute's Cancer Genome Atlas, many of the genes that have been proven to be involved in tumorigenesis
have been compiled into a single resource [52] . The Cancer Genome Atlas has assembled a list of 386 genes that impart tumorigenicity to the tissue. A privately funded group compiled a list of putative oncogenes, tumor suppressors and proto-oncogenes by using literature mining methods and called the resource the Cancer Genetics Web. The two resources were combined into single comprehensive list using a PERL script that eliminated redundancy of genes. The final list of potential cancer causing genes was uploaded into the MAPI resource using MySQL (Table 3-4).

Table 3-4: Genes involved in the progression of cancer

All genes shown to function as oncogenes, proto-oncogenes and tumor suppressors are listed in the cancer related genes table. The cytogenetic band of each cancer related gene is listed with its corresponding refseq ID.

Table 3-4: Genes involved in the progression of cancer

Cancer related genes
Alias
Gene symbol
Cytogenetic band
Refseq ID

Transcriptional regulation of genes

The National Center for Biotechnology Information hosts the Unigene database [53]. Unigene is a centralized, non-redundant database that organizes expressed sequence tags into gene oriented clusters. The Unigene database was downloaded from the FTP server at ftp://ftp.ncbi.nih.gov/repository/UniGene/. Human transcripts were extracted from the database using a PERL script.

There are over 120,000 unique transcripts expressed by the human genome. Unigene organizes the transcriptional information into the body site that expresses the EST, disease state, expression level of the transcript, and the developmental state. The unigene database lists each unique transcript and the number of times that transcript has been seen in that particular state. For the MAPI interface, the expression level of each transcript was transformed to represent the number of transcripts per one million. (Table 3-5).

Tissues accomplish their functional goals by regulated expression of necessary gene products. Each tissue expresses a unique number and combination of genes. For this study, we compared the number of unique transcripts between the normal prostate gland and the cancerous prostate gland (Figure 3-1). When the prostate gland becomes cancerous, the number of expressed gene products is significantly decreased. In addition to a decrease in number of the genes expressed, the average level of expression of each gene increases as compared to the normal gland (Table 3-6).

Table 3-5: Tissue Specific Gene expression levels

Each unique unigene transcript is listed, along with the developmental stage/ health state associated with each transcript and a normalized level of transcription.

Table 3-5: Tissue Specific Gene expression levels

Gene Expression Table
Gene ID in unigene format Type of transcription Tissue Expression level in transcripts per million

Table 3-6: Comparison of gene expression in the normal and cancerous prostate gland

Utilizing data obtained from Unigene, we determined that the number of unique RNA transcripts decreases in the cancerous prostate gland, as compared to the normal prostate. The tumorigenic tissue expresses approximately 30% fewer genes than the non-cancerous tissues. Even though the overall number of unique transcripts is fewer, the transcriptional level of each transcript is higher in the cancerous tissue than the non-cancerous tissue. At least in the prostate gland, when the gland becomes cancerous there is a significant increase in the expression level of a smaller number of genes.

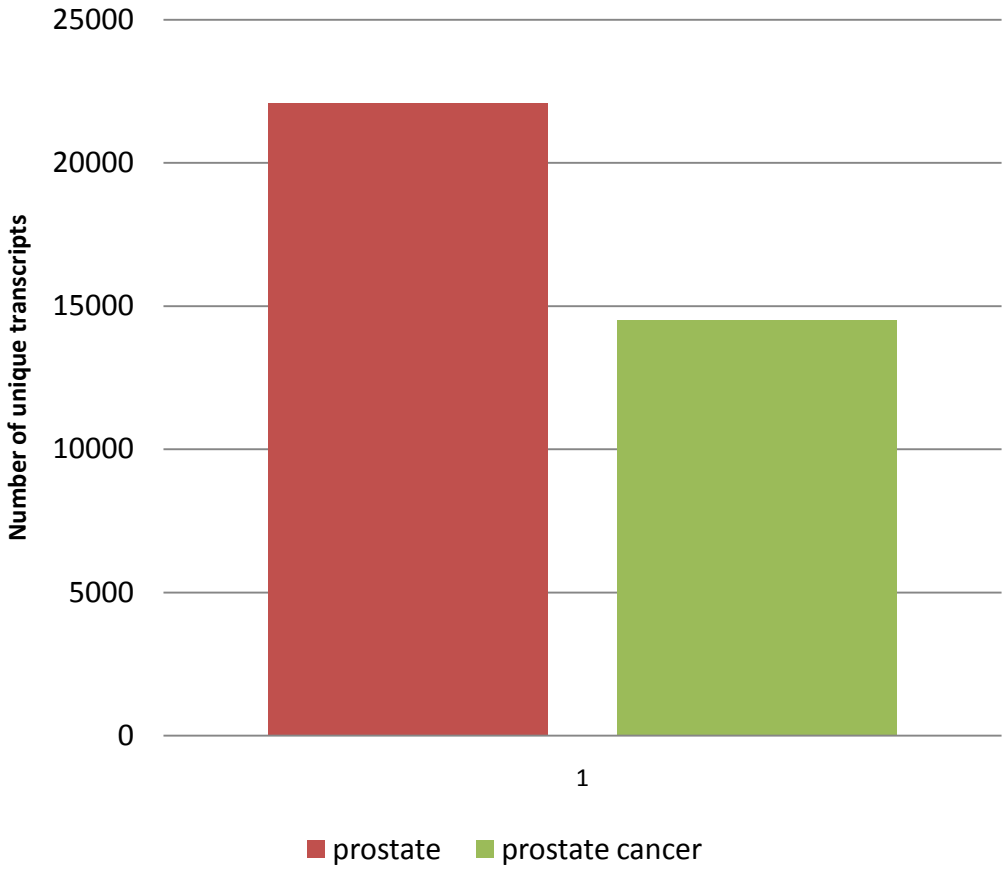
Table 3-6: Comparison of gene expression in the normal and cancerous prostate gland

Condition	Number of transcripts	Average expression	Minimum expression	Maximum expression
Cancer	14505	68.94	9.6	8040.8
Normal	22087	45.26	5.2	5974.8

Figure 3-1: Unique transcripts in the normal and cancerous prostate gland

The number of unique transcripts in the normal and cancerous prostate gland were extracted and compared to one another. The normal prostate gland expresses approximately 23,000 unique mRNA transcripts. When the gland becomes cancerous, the number of unique transcripts decreases to approximately 15,000. The prostate expresses approximately 15% of the total number of possible transcripts in the human genome.

Figure 3-1: Unique transcripts in the normal and cancerous prostate gland



Genes of the human genome

Human genes were extracted from a collection of chromosomal files contained in the Genbank resource at NCBI [54]. Genbank is a non-redundant collection of nucleotide sequences submitted by researchers in the biomedical field. Even though the resource is maintained by the US National Institute of Health, the database is synchronized daily with the European Molecular Biology Laboratory (EMBL) nucleotide sequence database and the DNA Databank of Japan. Genbank provides a comprehensive collection of all known genetic sequences. The database was downloaded on March 2, 2008 from ftp://ftp.ncbi.nih.gov/genbank/genomes/Eukaryotes/vertebrates_mammals/Homo_sapiens/GRCh37. A PERL script was written to parse desired information from the flat format sequence file. The information extracted from Genbank was used to create the genes table of the MAPI interface (Table 3-7).

Table 3-7: MAPI Human Genes Table

The genes table in the MAPI interface contains the reference sequence ID (Refseq), the gene symbol, chromosome of the gene, start and end of the gene and a short description of the role of the gene in the cell for all known human genes.

Table 3-7: MAPI Human Genes Table

Genes
Refseq ID
Contig ID
Chromosome
Chromosomal start
Chromosomal end
Direction
Gene symbol
Description

Tissue Specific MicroRNA Prediction

The potential benefit of tissue co-expression of microRNAs and their targets was evaluated, the predictions of Diana Micro T 3.0 and Pictar were overlapped and a set of targets from the union was generated. The list of predictions was compared to a list of genes that were expressed in the prostate gland. This combination of prediction tools was chosen because the union of the two datasets was proven to have the highest level of sensitivity. The tissue specific predicted target list was labeled as the MAPI dataset. If consideration is not given to tissue co-expression, the unioned dataset was shown to offer users a near 52% sensitivity but the number of predicted targets neared 400,000 (Table 2-1).

A plot similar to the plot in Figure 2-3 was created and amended to plot the sensitivity of each prediction tool against the average number of predicted targets per microRNA molecule. In order to measure the sensitivity of tissue specific target prediction, a subset of proven interactions that show an expression level of at least 1 transcript per million in the prostate gland was assembled and used as a comparison set. The methodology is similar to that described in chapter 2. MicroRNA prediction tools were evaluated using the standardized set of prostate specific interaction pairs. The sensitivity of nearly every prediction tool was increased with the exception of Miranda because proven interactions that are not expressed in the prostate were excluded from the comparison set (Figure 3-3). The sensitivity of Miranda actually decreased slightly when considering tissue specificity. The MAPI dataset, assembled from the union of PicTar and Diana MicroT 3.0 had an increased sensitivity and a lower number of predictions.

It appears that including tissue specification increases the sensitivity of prediction algorithms and concurrently increases specificity. The benefit to microRNA researchers is that

Figure 3-2: Comparison of microRNA prediction tools ranked by average number of predictions per microRNA

The sensitivity of each microRNA prediction tool was evaluated and plotted against the average number of predicted targets per microRNA molecule.

Figure 3-2: Comparison of microRNA prediction tools ranked by average number of predictions per microRNA

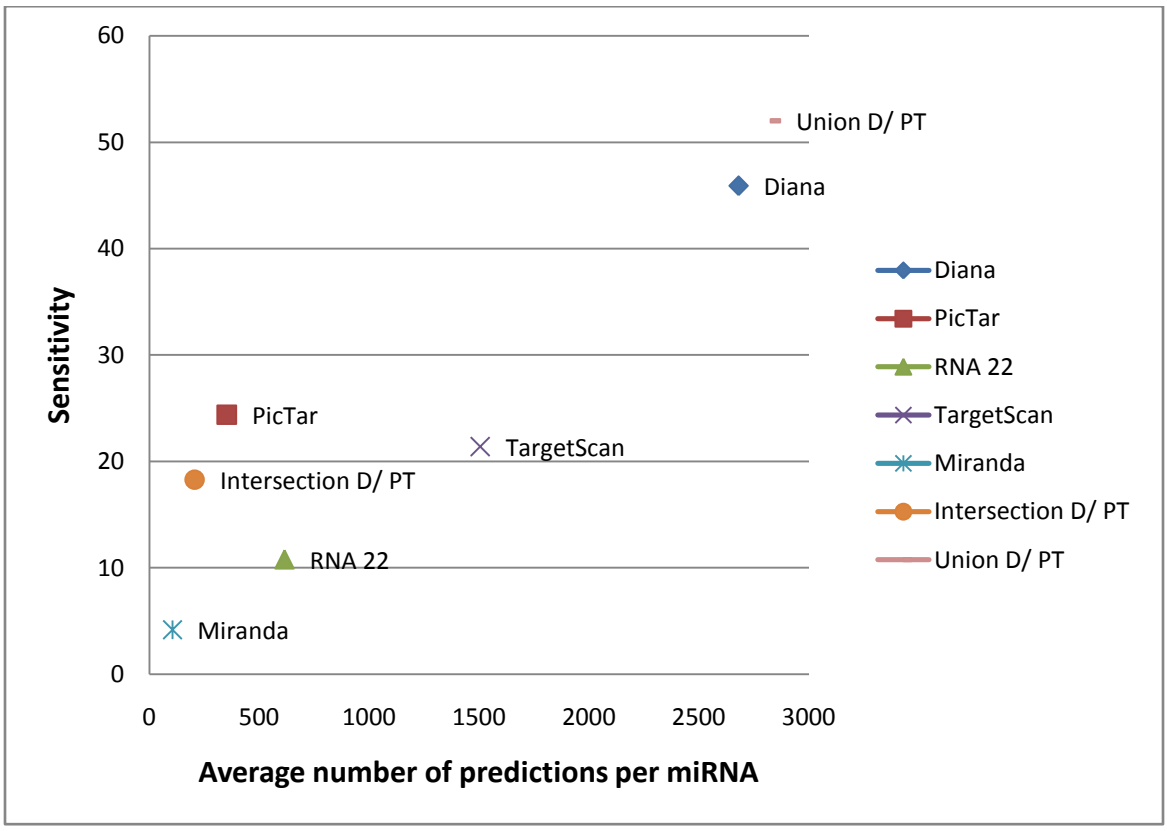
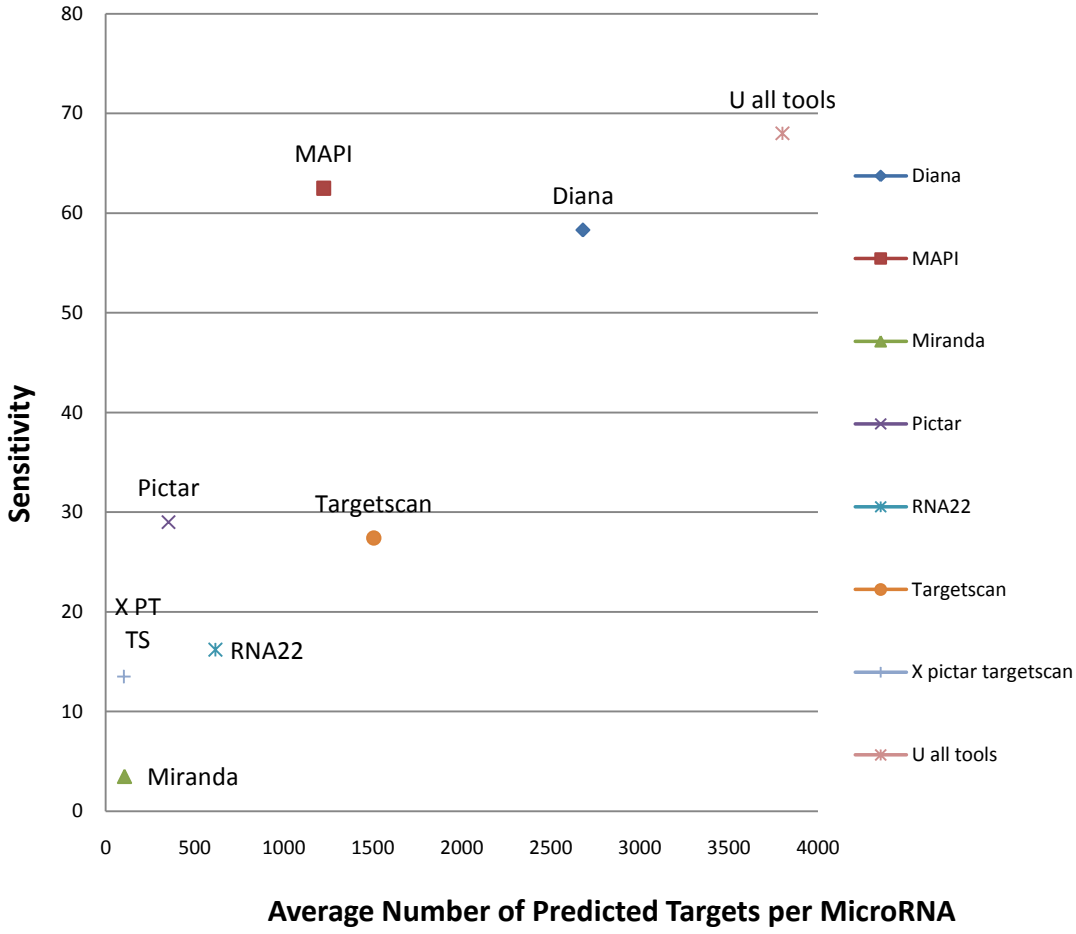


Figure 3-3: Comparison of microRNA prediction tools using tissue specific filtration

Considering tissue specific gene expression increases the sensitivity of all microRNA prediction tools and decreases the average number of predicted targets per microRNA molecule. MAPI predictions (union of Pictar and Diana Micro T 3.0) offer users the best balance between sensitivity and specificity. 62% of all proven interactions are predicted by this combination of prediction tools.

Figure 3-3: Comparison of microRNA prediction tools using tissue specific filtration



the end user is more likely to find an interaction if it exists and they will have a lower number of potential targets to evaluate.

The dataset labeled as MAPI had a sensitivity of over 60%. The only group of predictions that had a higher sensitivity was the dataset assembled by considering any target predicted by any of the prediction tools (union of all tools). MAPI was able to achieve this level of sensitivity, while only predicting 1000 possible targets per microRNA molecule. The dataset assembled from all prediction tools predicted an average of 3800 potential targets per microRNA molecule. The number of predicted targets per microRNA from MAPI may be more biologically relevant than the number achieved by the union of all prediction tools. Previous researchers have suggested that each microRNA may regulate up to 500 targets [12]. Inclusion of tissue specific transcriptional profiles increases the accuracy of computational prediction methods of microRNA target identification.

Chapter 4

Identification of Potential Targets of Human MicroRNA-17-3p Using MAPI

Identification of potential targets of HSA-miR-17-3p

It was discussed in chapter 1, that the level of human microRNA-17-3p varies in a cancer cell progression model and decreases as the cell becomes more tumorigenic. MicroRNA 17-p was more abundant in the normal, non-tumorigenic cell line (P69) and markedly decreased in the highly tumorigenic, metastatic cell line (M12) but increased in its weakly tumorigenic variant F6 [24]. It is also known that in patient tumor samples, normal epithelial tissue expresses higher levels of miR17-3p than tumor cells [21]. In fact, it was noted that as the Gleason score of cancer increases, there is a negative correlation to the level of miR-17-3p. That is as the cancer becomes less differentiated and more aggressive, the level of miR-17-3p declines. These observations prove that microRNA-17-3p functions as a tumor suppressor in the prostate gland.

Previous to the start of this work, it was shown that microRNA-17-3p regulates levels of vimentin, an intermediate filament protein [22]. It is hypothesized that most microRNAs regulate many targets. In order to determine other potential targets of miR-17-3p, we used MAPI described in Chapter 3. Priority was given to targets of human microRNA-17-3p that are expressed in the prostate gland and proven to be implicated in any form of human cancer (Table 4-1). Our search revealed two potential targets, insulin growth factor receptor 1 (*IGF1R*), and the Yamaguchi sarcoma viral oncogene homolog 1 (*YES1*). Table 4-2 describes the potential targets of microRNA-17-3p.

Table 4-1: MAPI Search Parameters for Tumorigenic Targets of microRNA-17-3p

Version 1.0 of the MicroRNA Annotation and Prediction Interface was queried to identify potential targets of the tumor suppressing microRNA-17-3p that are expressed in the prostate gland. The query options utilized are described in the table.

Table 4-1: MAPI Search Parameters for Tumorigenic Targets of microRNA-17-3p

Field	Parameter
MicroRNA	Human microRNA-17-3p
Tools	PicTar and Diana Micro T 3.0 union
Cancer related	Yes
Tissue	Prostate
Health State	Normal

Table 4-2: Potential Targets of Human microRNA-17-3p

Potential targets of miR17-3p expressed in the prostate gland and implicated in cancer. *YES1* is a member of the *src* family and possesses non-receptor tyrosine kinase activity [55]. *IGF1R* is a trans membrane receptor tyrosine kinase implicated in many forms of cancer [56].

Table 4-2: Potential Targets of Human MicroRNA-17-3p

Gene	Refseq ID	Chromosome/ cytogenetic band	Function
Yamaguchi Sarcoma Viral Homolog 1(Yes1)[13]	NM_005433	18p11	src family member/ non-receptor tyrosine kinase
Insulin Growth Factor Receptor 1 (IGF1R)[13]	NM_00875	15q26	Binds insulin growth factor and results in hypertrophy of cells

Structural Analysis of *IGF1R* and miR-17-3p Dimer

The 3'UTR of the insulin growth factor receptor gene and the sequence of human microRNA-17-3p were submitted to RNAhybrid, a tool for the determination of minimum free energy hybridization of long and short RNA molecules [48]. miR-17-3p and *IGF1R* are able to achieve dimerization at -31.7 kcal/mol. It is generally thought that true interactions between microRNA and target genes will have a delta G of less than -25.0 kcal/mol. The predicted structure has perfect seed region binding with bases 2-8 of the microRNA bound to the 3' UTR of *IGF1R*. The eighth position has a G:U wobble but all other bases of the seed participate in canonical Watson Crick base pair interactions. Two base pairs of the microRNA are unable to bind to the gene and loop out. There is significant 3' compensatory loop interactions of the microRNA. The predicted dimer structure appears to satisfy all of the rules identified for microRNA/ gene binding. (Figure 4-1a).

Multiple Sequence Alignment of *IGF1R* 3' UTRs

3' untranslated regions of the *IGF1R* gene were downloaded using the UCSC genome browser [57]. The sequences were compiled into a single fasta formatted file using PERL and submitted to the ClustalX program locally [58]. Figure 4-1b shows the regions of the alignment that correspond to the predicted binding site of *IGF1R*. It is generally thought that cross species conservation of nucleotides results from selective pressures to retain a given sequence and mutations in highly conserved regions are thought to be fatal. The majority of the 3' untranslated region of a gene is not conserved across multiple species, so regions of conservation have been

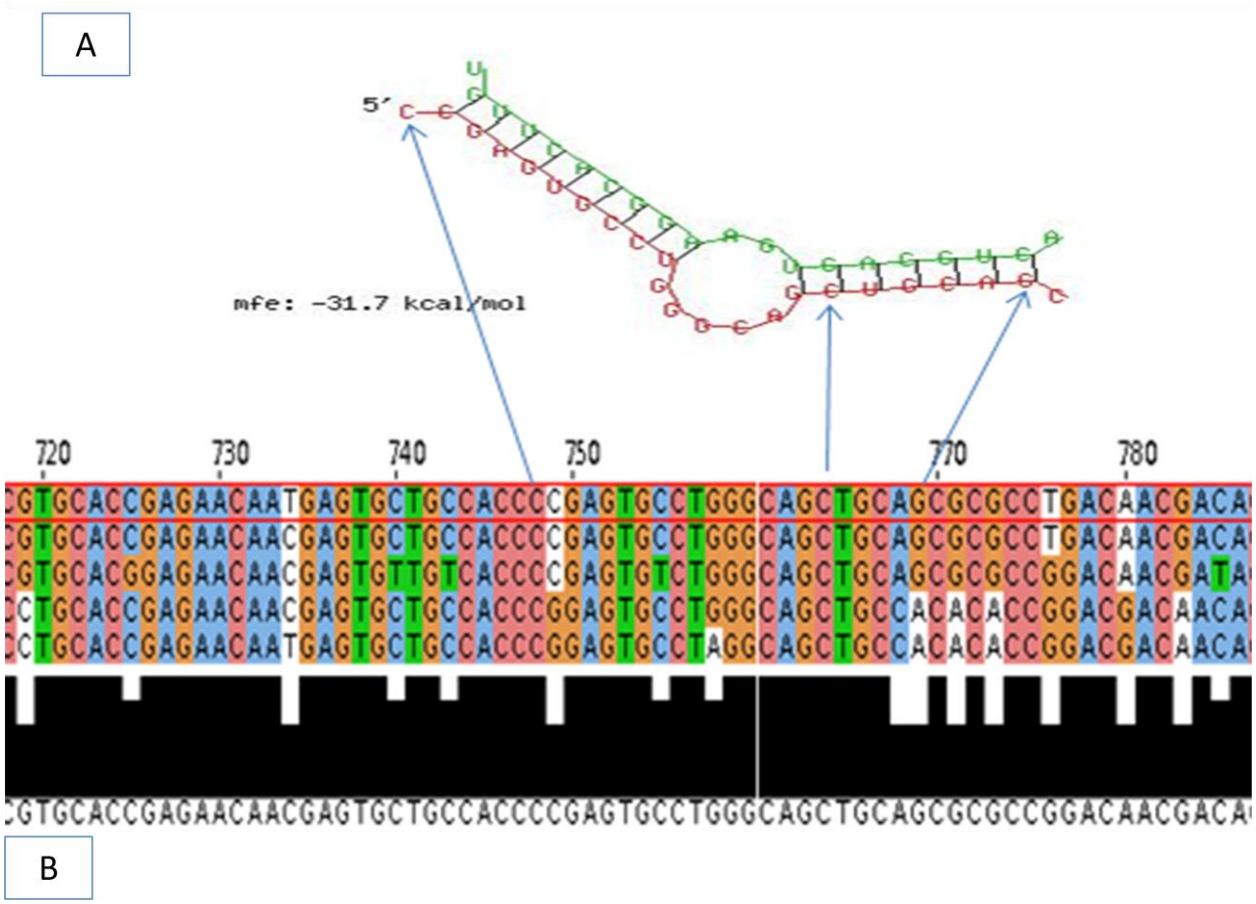
obviously retained for functional purposes. The predicted binding site of miR-17-3p is highly conserved across the species suggesting that this region of the 3' UTR may be involved in an interaction with microRNA-17-3p.

Figure 4-1: Multiple Sequence Alignment of *IGF1R* UTR and Predicted Structure

A. RNAhybrid predicted structure of microRNA-17-3p and the 3' UTR of insulin growth factor receptor gene [48]. The predicted minimum free energy is -31.7 kcal/ mol. The maximum free energy of a true target is generally thought to be -25.0 kcal/ mol. The overall structure adheres to all known "rules" observed in previous proven targets.

B. *IGF1R* sequences from human, mouse, rat, dog, and chimpanzee were aligned using Clustal X [58]. With the exception of the second base in mice and rats, all bases of the seed region are perfectly conserved across all species. Much of the 3' UTR not shown in this figure is not conserved. This implies a selective pressure for this region of the gene.

Figure 4-1: Multiple Sequence Alignment of *IGF1R* UTR and Predicted Structure



Structural Analysis of *YES1* and miR-17-3p

The sequence of the Yamaguchi Sarcoma Virus Oncogene Homolog was obtained from NCBI and submitted to the RNAhybrid program, along with the sequence of human microRNA-17-3p [48]. The predicted minimum free energy of dimer formation is -22.5 kcal/mol. Bases 2-8 of the seed region are predicted to bind to the 3' UTR of the gene, with a G:U wobble at position seven of the microRNA. There is a large loop of the UTR and a region of the microRNA that is not predicted to bind to the gene. Figure 4-2 shows the predicted structure of the miR and the UTR of the *YES1* gene. The predicted structure did not support the prediction of *YES1* as a potential target and it was not investigated further.

Insulin Growth Factor Receptor 1

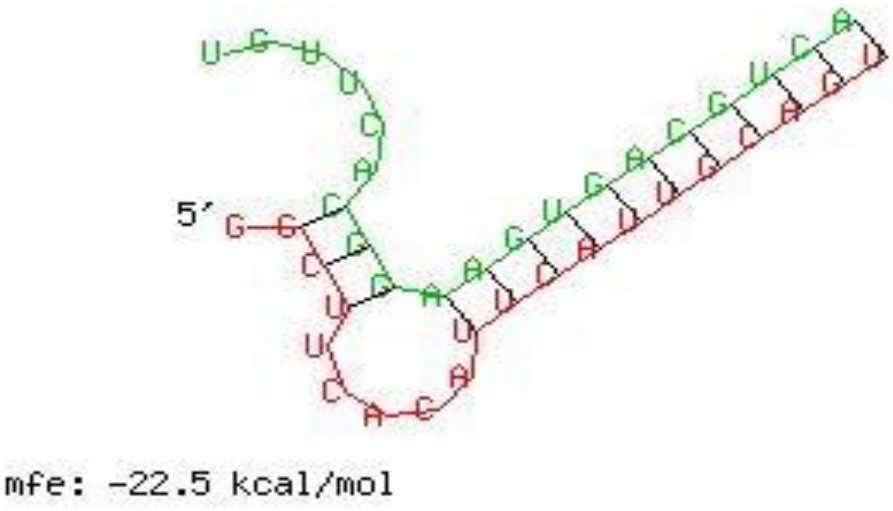
The insulin growth factor receptor 1 gene imparts a survival advantage, an anchorage independent growth advantage to cells and protects cells from apoptosis [56, 59]. Cells that express higher levels of *IGF1R* are shown to be more tumorigenic and metastatic. It is hypothesized that increased levels of *IGF1R* lead to a transformed phenotype that imparts an oncogenic potential to the cell.

Tissue samples extracted from formalin-fixed, paraffin embedded tissue obtained from a fine needle biopsy showed higher protein and mRNA levels of *IGF1R* in primary prostate cancer compared to benign prostatic epithelium [59]. Levels of insulin growth factor receptor 1 were also shown to increase in the metastatic sites of prostate cancer. Metastasis is accomplished because metastasizing cells must detach themselves from other cells in the tissue and mobilize

Figure 4-2: Predicted Structure of miR-17-3p and *YES1*

RNAhybrid predicted structure of dimer formation between microRNA-17-3p and the 3' UTR of *YES1* [48].

Figure 4-2: Predicted Structure of miR-17-3p and *YES1*



themselves. Vimentin has been shown to correlate with an increased ability of cells to be motile and invasive [60]. *IGF1R* potentially causes an upregulation of extracellular proteases in the prostate, imparting an ability of cells to detach from their neighbors [61]. Because of the structural prediction, sequence conservation and functional significance, further validation of *IGF1R* was undertaken.

Validation of miR-17-3p and *IGF1R* Interaction

In order to validate the regulation of insulin growth factor receptor 1 levels in the prostate, protein levels were measured in a prostate cancer cell line [24]. Two sublines were chosen for comparison of protein levels. The first is a highly tumorigenic, metastatic cell line stably transformed with a plasmid that expresses a non-targeting RNA molecule (M12 +NTC). This cell line serves as our negative control. A second set of highly tumorigenic, metastatic cells was stably transformed with a plasmid that expresses a functional copy of microRNA-17-3p (M12 + miR17-3p) [22]. This cell line is our experimental model.

Cell Culture Methods

All cells were grown at 37° C in RPMI 1640 growth media containing L-glutamine obtained from Sigma-Aldrich and supplemented with 5% fetal bovine serum, 5 μg/ml insulin, 5 μg/ml transferring, 5 ng/ml selenious acid, (ITS from Collaborative Research Bedford,MA). Gentamycin (0.05 mg/ml) was added to inhibit bacterial contamination of culture. Cells containing the integrated plasmids were selected by growth in puramycin. All tissue culture cells

were grown in 250 ml T75 flasks and split when confluent. Cells were pelleted after trypsin digestion by centrifugation at 5000 rpm for five minutes. Cells pellets were washed in 1X PBS buffer and re-centrifuged at 5000 rpm for five additional minutes. Following pelleting of the cells, cells were flash frozen in liquid nitrogen and stored at -80 ° C.

Western Blot Analysis

Cell pellets were thawed on ice and re-suspended in 200 - 400 microliters of 4% SDS in PBS after thawing. Cell lysates were prepared by sonication of cell suspension for five minutes. Following sonication, cellular debris was removed after dilution of lysate in one volume of PBS buffer and centrifugation at 10,000 rpm for five additional minutes. Proteins (50 mg) were separated by electrophoresis on a BioRad SDS denaturing 4-14% Tris-HCl gradient gel at 120 mV for 1.5 hours.

Separated proteins were transferred to a nitrocellulose membrane and non-specific interactions were minimized by blocking the membrane in 3% powdered milk dissolved in TBST buffer. IGF1R proteins were visualized using IGF-I receptor beta antibody from Cell Signaling Technology and a secondary anti-rabbit antibody.

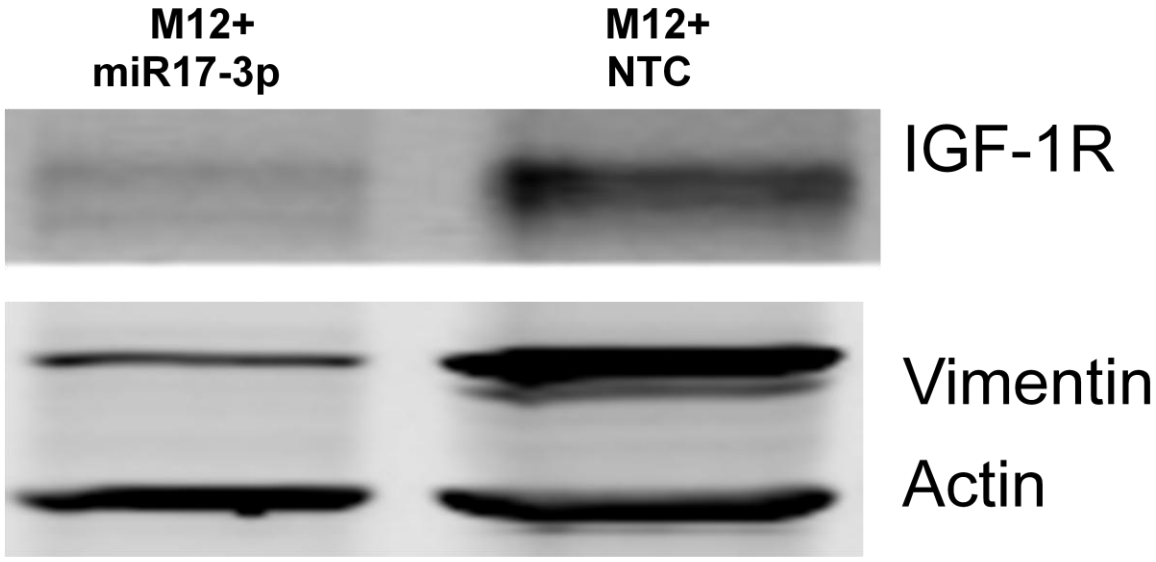
Figure 4-3 shows the results of the western blot analysis of the M12 + NTC and M12 + miR-17-3p cell lines. Actin was used as an internal control to verify consistency of protein loads. Vimentin levels were also analyzed to ensure that the plasmid transcribing miR-17-3p was functioning, as it is known that miR-17-3p targets vimentin. Protein levels of IGF1R were significantly higher in the M12 + NTC cell line as compared to the levels in the M12 + miR-17-3p cell line. The level of actin was consistent across both cell lines and the levels of vimentin exhibited the expected decrease in expression confirming the increased level of miR-17-3p in the

stably transformed M12+miR-17-3p cell line. The decrease in IGF1R levels dependent on miR-17-3p confirms that IGF1R mRNA is targeted by miR-17-3p.

Figure 4-3: Western Blot Analysis of IGF1R Protein Levels

Whole cell extracts were subjected to Western blot analysis with antibody to IGF1R, actin, or vimentin in M12 cells with restored expression of miR-17-3p (M12+ miR-17-3p) compared to M12 cells expressing a negative non-targeting RNA control (M12+NTC). Levles of actin are consistent in both lanes of the western blot, proving that the load of protein is similar in both lanes of the gradient gel. Vimentin levels were analyzed as a positive control, as it has been previously shown that miR-17-3p targets vimentin [22].

Figure 4-3: Western Blot Analysis of IGF1R Protein Levels



Conclusions

Prostate cancer is a significant problem for men in the United States and across the world. MicroRNA proteins are post-transcriptional regulators of protein products that have been shown to be involved in numerous cellular processes. MicroRNA dysregulation has been shown to lead to the development of several forms of human cancer. Human microRNA-17-3p has been shown to be differentially regulated in primary tumors of the prostate and decreases as the Gleason score of the tumor increases.

miR-17-3p has been shown to target the intermediate filament protein vimentin. It is hypothesized that most microRNAs target more than one protein. A bioinformatics approach was undertaken to elucidate further putative targets of miR-17-3p. A unique comprehensive microRNA target prediction tool was designed that harnesses the information inherent in many freely available databases and combines them into a single resource. This comprehensive database was used to identify other potential targets of miR-17-3p. *IGF1R* was identified as a potential target of the microRNA, which was previously shown to be differentially regulated in prostate cancer. Prostate cancer cell lines were utilized to verify regulation of *IGF1R* by miR-17-3p. Levels of IGF1R protein varied between the cell lines and was lower in the cell line expressing microRNA-17-3p. Based on these experiments, it does indeed appear as if microRNA-17-3p regulated the levels of insulin growth factor receptor and the MAPI interface was instrumental in identifying this new relevant target for miR-17-3p.

References

1. [http://www.cancer.gov/aboutnci/servingpeople/snapshots/prostate.pdf]

2. Andriole GL, Crawford ED, Grubb RL,3rd, Buys SS, Chia D, Church TR, Fouad MN, Gelmann EP, Kvale PA, Reding DJ, Weissfeld JL, Yokochi LA, O'Brien B, Clapp JD, Rathmell JM, Riley TL, Hayes RB, Kramer BS, Izmirlian G, Miller AB, Pinsky PF, Prorok PC, Gohagan JK, Berg CD, PLCO Project Team: **Mortality results from a randomized prostate-cancer screening trial.** N Engl J Med 2009, **360**(13):1310-1319.

3. Hsing AW, Chokkalingam AP: **Prostate cancer epidemiology.** Front Biosci 2006, **11**:1388-1413.

4. [http://apps.nccd.cdc.gov/uscs/]

5. Hankey BF, Feuer EJ, Clegg LX, Hayes RB, Legler JM, Prorok PC, Ries LA, Merrill RM, Kaplan RS: **Cancer surveillance series: interpreting trends in prostate cancer--part I: Evidence of the effects of screening in recent prostate cancer incidence, mortality, and survival rates.** J Natl Cancer Inst 1999, **91**(12):1017-1024.

6. Catalona WJ, Richie JP, Ahmann FR, Hudson MA, Scardino PT, Flanigan RC, deKernion JB, Ratliff TL, Kavoussi LR, Dalkin BL: **Comparison of digital rectal examination and serum prostate specific antigen in the early detection of prostate cancer: results of a multicenter clinical trial of 6,630 men.** J Urol 1994, **151**(5):1283-1290.

7. Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, Tamayo P, Renshaw AA, D'Amico AV, Richie JP, Lander ES, Loda M, Kantoff PW, Golub TR, Sellers WR: **Gene expression correlates of clinical prostate cancer behavior.** Cancer Cell 2002, **1**(2):203-209.

8. Taplin ME, Bubley GJ, Shuster TD, Frantz ME, Spooner AE, Ogata GK, Keer HN, Balk SP: **Mutation of the androgen-receptor gene in metastatic androgen-independent prostate cancer.** N Engl J Med 1995, **332**(21):1393-1398.

9. Dhanasekaran SM, Barrette TR, Ghosh D, Shah R, Varambally S, Kurachi K, Pienta KJ, Rubin MA, Chinnaiyan AM: **Delineation of prognostic biomarkers in prostate cancer.** Nature 2001, **412**(6849):822-826.

10. Lee RC, Feinbaum RL, Ambros V: **The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14.** Cell 1993, **75**(5):843-854.

11. Hoffmann R, Valencia A: **A gene network for navigating the literature.** Nat Genet 2004, **36**(7):664.

12. Shomron N, Golan D, Hornstein E: **An evolutionary perspective of animal microRNAs and their targets.** J Biomed Biotechnol 2009, **2009**:594738.

13. Safran M, Solomon I, Shmueli O, Lapidot M, Shen-Orr S, Adato A, Ben-Dor U, Esterman N, Rosen N, Peter I, Olender T, Chalifa-Caspi V, Lancet D: **GeneCards 2002: towards a complete, object-oriented, human gene compendium.** Bioinformatics 2002, **18**(11):1542-1543.

14. Shahi P, Loukianiouk S, Bohne-Lang A, Kenzelmann M, Kuffer S, Maertens S, Eils R, Grone HJ, Gretz N, Brors B: **Argonaute--a database for gene regulation by mammalian microRNAs.** Nucleic Acids Res 2006, **34**(Database issue):D115-8.

15. Alexiou P, Maragkakis M, Papadopoulos GL, Reczko M, Hatzigeorgiou AG: **Lost in translation: an assessment and perspective for computational microRNA target identification.** Bioinformatics 2009, **25**(23):3049-3055.

16. Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ: **miRBase: tools for microRNA genomics.** Nucleic Acids Res 2008, **36**(Database issue):D154-8.

17. Cui Q, Yu Z, Purisima EO, Wang E: **Principles of microRNA regulation of a human cellular signaling network.** Mol Syst Biol 2006, **2**:46.

18. Maragkakis M, Alexiou P, Papadopoulos GL, Reczko M, Dalamagas T, Giannopoulos G, Goumas G, Koukis E, Kourtis K, Simossis VA, Sethupathy P, Vergoulis T, Koziris N, Sellis T, Tsanakas P, Hatzigeorgiou AG: **Accurate microRNA target prediction correlates with protein repression levels.** BMC Bioinformatics 2009, **10**:295.

19. Calin GA, Croce CM: **MicroRNA-cancer connection: the beginning of a new tale.** Cancer Res 2006, **66**(15):7390-7394.

20. Esquela-Kerscher A, Slack FJ: **Oncomirs - microRNAs with a role in cancer.** Nat Rev Cancer 2006, **6**(4):259-269.

21. Zhang X, Ladd A, Dragoescu E, Budd WT, Ware JL, Zehner ZE: **MicroRNA-17-3p is a prostate tumor suppressor in vitro and in vivo, and is decreased in high grade prostate tumors analyzed by laser capture microdissection.** Clin Exp Metastasis 2009, .

22. Zhang X, Fournier M, Ware JL, Bissel MJ, Yacoub A, Zehner ZE: **Inhibition of vimentin or ß1-integrin reverts morphology of prostate tumor cells grown in laminin-rich extracellular matrix gels and reduces tumor growth *in vivo*.** Mol Can Ther 2009, **8**(3).

23. Goldman RD, Khuon S, Chou YH, Opal P, Steinert PM: **The function of intermediate filaments in cell shape and cytoskeletal integrity.** J Cell Biol 1996, **134**(4):971-983.

24. Bae VL, Jackson-Cook CK, Maygarden SJ, Plymate SR, Chen J, Ware JL: **Metastatic sublines of an SV40 large T antigen immortalized human prostate epithelial cell line.** Prostate 1998, **34**(4):275-282.

25. Astbury C, Jackson-Cook CK, Culp SH, Paisley TE, Ware JL: **Suppression of tumorigenicity in the human prostate cancer cell line M12 via microcell-mediated restoration of chromosome 19.** Genes Chromosomes Cancer 2001, **31**:143-155.

26. Singh S, Sadacharan S, Su S, Belldegrun A, Persad S, Singh G: **Overexpression of vimentin: role in the invasive phenotype in an androgen-independent model of prostate cancer.** Cancer Res 2003, **63**(9):2306-2311.

27. Baskerville S, Bartel DP: **Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes.** RNA 2005, **11**(3):241-247.

28. Lee Y, Jeon K, Lee JT, Kim S, Kim VN: **MicroRNA maturation: stepwise processing and subcellular localization.** EMBO J 2002, **21**(17):4663-4670.

29. Lee Y, Ahn C, Han J, Choi H, Kim J, Yim J, Lee J, Provost P, Radmark O, Kim S, Kim VN: **The nuclear RNase III Drosha initiates microRNA processing.** Nature 2003, **425**(6956):415-419.

30. Zeng Y, Cullen BR: **Structural requirements for pre-microRNA binding and nuclear export by Exportin 5.** Nucleic Acids Res 2004, **32**(16):4776-4785.

31. Hutvagner G, Zamore PD: **A microRNA in a multiple-turnover RNAi enzyme complex.** Science 2002, **297**(5589):2056-2060.

32. Lee Y, Jeon K, Lee JT, Kim S, Kim VN: **MicroRNA maturation: stepwise processing and subcellular localization.** EMBO J 2002, **21**(17):4663-4670.

33. Brennecke J, Stark A, Russell RB, Cohen SM: **Principles of microRNA-target recognition.** PLoS Biol 2005, **3**(3):e85.

34. Maziere P, Enright AJ: **Prediction of microRNA targets.** Drug Discov Today 2007, **12**(11-12):452-458.

35. Maziere P, Enright AJ: **Prediction of microRNA targets.** Drug Discov Today 2007, **12**(11-12):452-458.

36. John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS: **Human MicroRNA targets.** PLoS Biol 2004, **2**(11):e363.

37. Krek A, Grun D, Poy MN, Wolf R, Rosenberg L, Epstein EJ, MacMenamin P, da Piedade I, Gunsalus KC, Stoffel M, Rajewsky N: **Combinatorial microRNA target predictions.** Nat Genet 2005, **37**(5):495-500.

38. Lewis BP, Burge CB, Bartel DP: **Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets.** Cell 2005, **120**(1):15-20.

39. Miranda KC, Huynh T, Tay Y, Ang YS, Tam WL, Thomson AM, Lim B, Rigoutsos I: **A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes.** Cell 2006, **126**(6):1203-1217.

40. Hofacker IL: **RNA secondary structure analysis using the Vienna RNA package.** Curr Protoc Bioinformatics 2004, **Chapter 12**:Unit 12.2.

41. Hofacker IL: **RNA secondary structure analysis using the Vienna RNA package.** Curr Protoc Bioinformatics 2004, **Chapter 12**:Unit 12.2.

42. Rigoutsos I, Floratos A, Ouzounis C, Gao Y, Parida L: **Dictionary building via unsupervised hierarchical motif discovery in the sequence space of natural proteins.** Proteins 1999, **37**(2):264-277.

43. Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A: **Rfam: annotating non-coding RNAs in complete genomes.** Nucleic Acids Res 2005, **33**(Database issue):D121-4.

44. Betel D, Wilson M, Gabow A, Marks DS, Sander C: **The microRNA.org resource: targets and expression.** Nucleic Acids Res 2008, **36**(Database issue):D149-53.

45. Sethupathy P, Megraw M, Hatzigeorgiou AG: **A guide through present computational approaches for the identification of mammalian microRNA targets.** Nat Methods 2006, **3**(11):881-886.

46. Xiao F, Zuo Z, Cai G, Kang S, Gao X, Li T: **miRecords: an integrated resource for microRNA-target interactions.** Nucleic Acids Res 2009, **37**(Database issue):D105-10.

47. Papadopoulos GL, Reczko M, Simossis VA, Sethupathy P, Hatzigeorgiou AG: **The database of experimentally supported targets: a functional update of TarBase.** Nucleic Acids Res 2009, **37**(Database issue):D155-8.

48. Rehmsmeier M, Steffen P, Hochsmann M, Giegerich R: **Fast and effective prediction of microRNA/target duplexes.** RNA 2004, **10**(10):1507-1517.

49. Ritchie W, Flamant S, Rasko JE: **Predicting microRNA targets and functions: traps for the unwary.** Nat Methods 2009, **6**(6):397-398.

50. Chung CH, Bernard PS, Perou CM: **Molecular portraits and the family tree of cancer.** Nat Genet 2002, **32 Suppl**:533-540.

51. Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, Sweet-Cordero A, Ebert BL, Mak RH, Ferrando AA, Downing JR, Jacks T, Horvitz HR, Golub TR: **MicroRNA expression profiles classify human cancers.** Nature 2005, **435**(7043):834-838.

52. Cancer Genome Atlas Research Network: **Comprehensive genomic characterization defines human glioblastoma genes and core pathways.** Nature 2008, **455**(7216):1061-1068.

53. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Landsman D, Lipman DJ, Lu Z, Madden TL, Madej T, Maglott DR, Marchler-Bauer A, Miller V, Mizrachi I, Ostell J, Panchenko A, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Shumway M, Sirotkin K, Slotta D, Souvorov A, Starchenko G, Tatusova TA, Wagner L, Wang Y, John Wilbur W, Yaschenko E, Ye J: **Database resources of the National Center for Biotechnology Information.** Nucleic Acids Res 2010, **38**(Database issue):D5-16.

54. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL: **GenBank.** Nucleic Acids Res 2008, **36**(Database issue):D25-30.

55. Brickell PM: **The p60c-src family of protein-tyrosine kinases: structure, regulation, and function.** Crit Rev Oncog 1992, **3**(4):401-446.

56. Mitsiades CS, Mitsiades NS, McMullan CJ, Poulaki V, Shringarpure R, Akiyama M, Hideshima T, Chauhan D, Joseph M, Libermann TA, Garcia-Echeverria C, Pearson MA, Hofmann F, Anderson KC, Kung AL: **Inhibition of the insulin-like growth factor receptor-1 tyrosine kinase activity as a therapeutic strategy for multiple myeloma, other hematologic malignancies, and solid tumors.** Cancer Cell 2004, **5**(3):221-230.

57. Rhead B, Karolchik D, Kuhn RM, Hinrichs AS, Zweig AS, Fujita PA, Diekhans M, Smith KE, Rosenbloom KR, Raney BJ, Pohl A, Pheasant M, Meyer LR, Learned K, Hsu F, Hillman-Jackson J, Harte RA, Giardine B, Dreszer TR, Clawson H, Barber GP, Haussler D, Kent WJ: **The UCSC Genome Browser database: update 2010.** Nucleic Acids Res 2010, **38**(Database issue):D613-9.

58. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG: **Clustal W and Clustal X version 2.0.** Bioinformatics 2007, **23**(21):2947-2948.

59. Hellawell GO, Turner GD, Davies DR, Poulsom R, Brewster SF, Macaulay VM: **Expression of the type 1 insulin-like growth factor receptor is up-regulated in primary prostate cancer and commonly persists in metastatic disease.** Cancer Res 2002, **62**(10):2942-2950.

60. Zhao Y, Yan Q, Long X, Chen X, Wang Y: **Vimentin affects the mobility and invasiveness of prostate cancer cells.** Cell Biochem Funct 2008, **26**(5):571-577.

61. Saikali Z, Setya H, Singh G, Persad S: **Role of IGF-1/IGF-1R in regulation of invasion in DU145 prostate cancer cells.** Cancer Cell Int 2008, **8**:10.

Vita

William Thomas Budd was born August 27, 1973 in Alexanderia, Virginia and is a citizen of the United States of America. He graduated in 1991 from Orange County High School in Orange, Virginia. He graduated Magna Cum Laude and received his Bachelors of Science in Bioinformatics from Virginia Commonwealth University in 2009. He has taught Introduction to Life Sciences and Application in Bioinformatics at Virginia Commonwealth University during the pursuit of his graduate degree. He has also taught Anatomy and Physiology at the Medical Careers Institute of the ECPI School of Health Science. Publications include a December 2009 manuscript published in Clinical and Experimental Metastasis entitled "MicroRNA-17-3p is a Prostate Tumor Suppressor *In Vitro* and *In Vivo*, and is Decreased in High Grade Prostate Tumors Analyzed by Laser Capture Microdissection".